INFO SERVICES
Transformation Excellence

Whitepaper

Scalable Cloud Data Lake

Author : Raghavan Madabusi

Introduction

In today's digital age, the sheer volume of data generated and consumed by businesses is staggering. As organizations grapple with the challenges of harnessing, analyzing, and deriving value from this data, traditional data storage and processing methods often fall short.

This whitepaper delves into the intricacies of different cloud data lake architecture patterns, exploring its benefits, components, the transformative potential it holds for businesses, and covers AWS's Modern Data Architecture that enables businesses to rapidly build scalable cloud data lakes.

Cloud Data Lake Architecture: A New Paradigm

To truly grasp the transformative potential of cloud data lakes, it's essential to journey back in time and understand the traditional data processing landscape. Historically, businesses viewed data as a secondary element, a tool to address specific business challenges.

The problem-first approach typically followed these steps:

Pinpoint a specific business challenge.

Design a data structure tailored to address this challenge.

Gather or produce data that aligns with this predefined structure.

Store this structured data in an OLTP system, like Oracle Database.

Execute additional transformations (such as sorting, grouping, etc.) to transfer data into OLAP systems; Oracle databases often played a role here too.

Construct visualizations and queries from these OLAP systems to tackle the business challenge.

Limitations of the Problem-First Approach:

Limited Scope

By focusing solely on the immediate problem, businesses might overlook other valuable insights that their data can offer.

Inflexibility

Predefined data structures can become restrictive, especially when new types of data or new business challenges emerge.

Scalability Issues

As data volumes grow, traditional OLTP and OLAP systems might struggle to handle the influx, leading to performance bottlenecks.

Integration Challenges

With data siloed in different databases tailored to specific problems, achieving a unified view becomes complex and resource-intensive.

Costly Overheads

Any change or evolution in the business problem might necessitate restructuring the data, leading to increased costs and time delays.

Understanding Cloud Data Lake Architecture

Unlike traditional systems that were often constrained by data source, format, or size, cloud data lakes are inherently flexible. They prioritize a data-centric approach over a problem-centric one. In this paradigm, every piece of data, regardless of its current relevance, is seen as potentially valuable.

In the cloud data lake architecture, data is ingested in its most authentic form, free from limitations related to its origin or structure. This data finds its home in the cloud data lake, a versatile storage system designed for diverse data types. While the raw data can vary in quality and significance, it undergoes transformations to yield meaningful insights.

Advantages of Cloud Data Lake Architecture

Cloud data lake architectures offer several distinct advantages over traditional data warehouse systems:

01

Unrestricted Data Handling

Cloud data lakes are equipped to manage any data type, irrespective of its source, volume, or structure. They can seamlessly handle both real-time data streams and batch data. Moreover, the cost-effective nature of data lake storage ensures that data retention doesn't strain the budget.

02

Unified Storage System

One of the standout features of a cloud data lake is its singular storage layer, eliminating the need for multiple specialized storage systems. This not only reduces costs but also minimizes potential errors that can arise during data transfers.

03

Diverse Processing Capabilities

The decoupled nature of compute and storage in a cloud data lake allows for a range of data processing tools to operate on a single storage layer. Whether it's BI analytics, advanced machine learning, or domain-specific computations, the architecture supports it all.

04

Cost-Effective Scalability

Cloud tools are designed for elasticity. They can be scaled up or down based on demand, ensuring you only pay for what you use. This is especially beneficial during peak demand periods, allowing for resource optimization without long-term commitments.

05

Flexible Resource Allocation

The architecture's ability to scale compute and storage resources independently is a game-changer. While storage on the cloud is cost-effective, compute resources, though pricier, can be dynamically managed, ensuring optimal resource utilization.

Cloud Data Lake Architecture Patterns:

1

Modern Data Warehouse Architecture

In a modern data warehouse architecture, both the data lake and the data warehouse peacefully coexist, each serving a distinct purpose.

2

Data Lakehouse Architecture

A data lakehouse is a data platform, which merges the best aspects of data warehouses and data lakes into one data management solution.

3

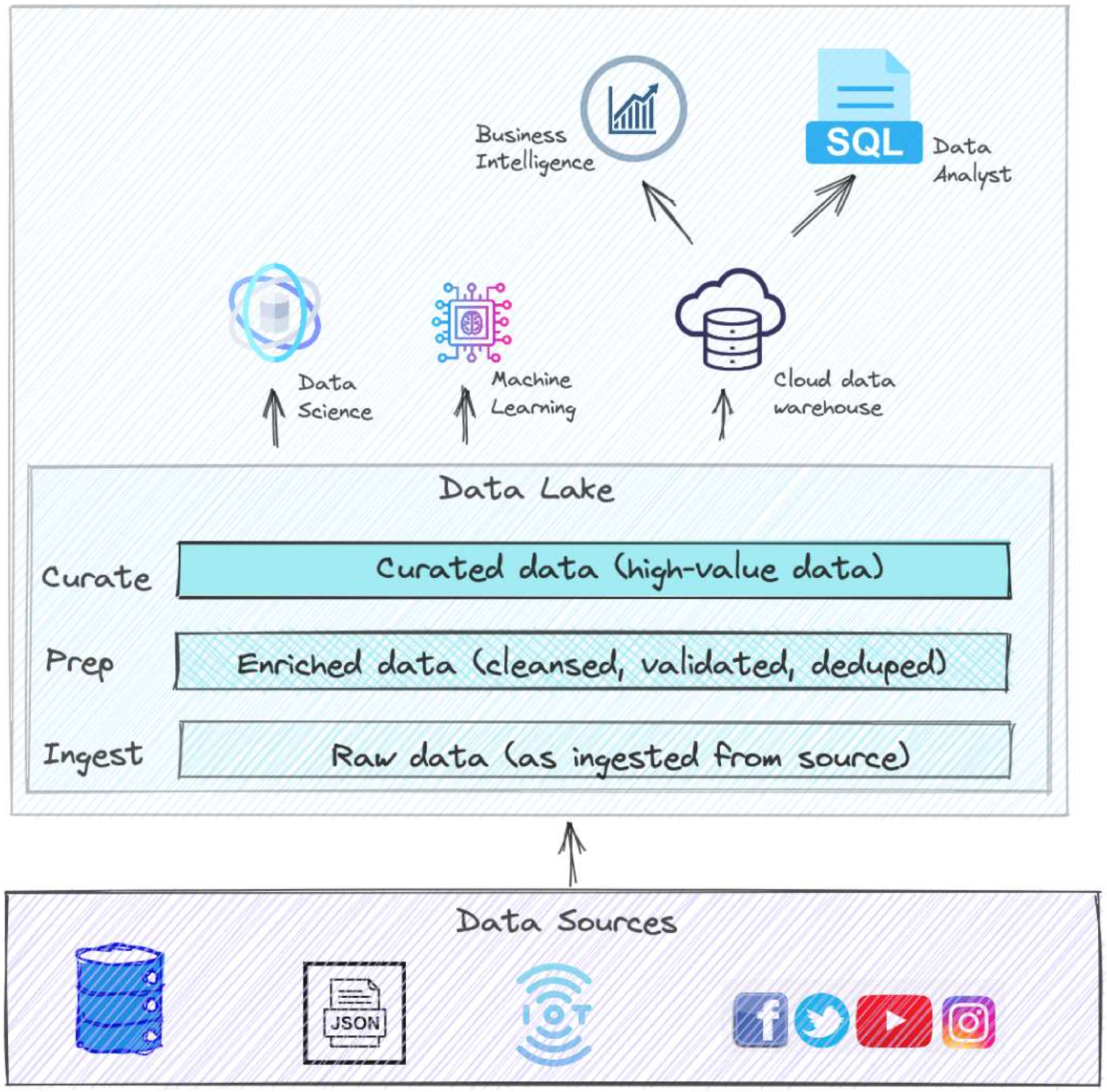
Data Mesh Architecture

Data mesh architectures shift the focus from a centralized data infrastructure to domain-specific data products. They embrace the concept of data as a product in each domain and require domains and data products to adhere to interoperability standards.

Modern Data Warehouse Architecture

Within the framework of a modern data warehouse architecture, the data lake and data warehouse harmoniously coexist, each with its unique role. The data lake acts as an economical storage solution for vast data quantities, facilitating advanced tasks like data science and machine learning. On the other hand, the data warehouse retains high-priority data, driving business dashboards and offering insights through structured data analysis.

Architectural Overview Data, sourced from diverse origins like in-house databases or social media streams, is initially directed into a data lake. Subsequent transformations are carried out using advanced analytics tools such as Hadoop and Spark. These tools enable the consolidation and refinement of datasets into structured, high-value data. This refined data is then channeled into a cloud-based data warehouse, which fuels various dashboards, including dynamic ones for BI professionals who prefer SQL-based tools. Additionally, the data lake paves the way for innovative analyses by data experts and the development of machine learning algorithms that can be integrated into applications.



The rationale for incorporating a data lake into this framework includes:

01

Data lakes are more cost-effective than data warehouses, serving as a sustainable data reservoir. Given the vast data volumes typically stored in data lakes, the cost difference is significant.

02

Data lakes are compatible with a plethora of contemporary data science and machine learning tools, unlocking novel analytical possibilities.

03

Data lakes offer a scalable design that can adapt to evolving requirements. For instance, an initial data lake setup might cater to periodic data uploads from in-house systems and report generation. However, this setup can be expanded to accommodate real-time data feeds without a complete overhaul.

04

Organizations are increasingly recognizing the value of diverse data types. Even if the current focus is on structured data, there's potential value in diverse data sources like meteorological data or social media activity.

Advantages of Modern Data Warehouse Architecture

01 Empowers business analysts to use familiar BI tools, primarily SQL-based.

02 Unlocks new analytical avenues, such as data science and machine learning.

03 Incorporates a data lake, enhancing storage and analytical capabilities.

04 Retains conventional data warehouse features, like SQL interfaces for BI professionals.

05 Allows data access restriction to the data warehouse using established access control techniques.

06 Facilitates the transition of on-site applications to the cloud, reducing dual infrastructure maintenance.

07 Offers cost savings by archiving operational data in a data lake for extended periods.

Limitations of Modern Data Warehouse Architecture

01 Requires managing two distinct infrastructures: the data lake and the data warehouse.

02 Ensuring data quality in the versatile environment of a data lake can be challenging.

03 Risk of the data lake turning into a "data swamp" without proper management.

04 Dependency on data engineers to process and transfer new data sets to the warehouse.

05 Potential bottlenecks when BI professionals require new datasets.

06 Need to revert valuable data segments from the warehouse to the data lake for exploratory analysis, complicating data sharing.

Data Lakehouse Architecture

The term "lakehouse" in the data realm, popularized by Databricks, has become a significant industry buzz. Amazon initially coined the term "lake house" (with a space) when launching Redshift Spectrum. The concept gained traction in January 2020, thanks to a Databricks article that described the lakehouse as a fusion of data lakes and data warehouses' best features.

The lakehouse model essentially combines

Data lakes for analytical, data science, and machine learning applications.

Data warehouses for SQL-based interactive queries and BI applications.

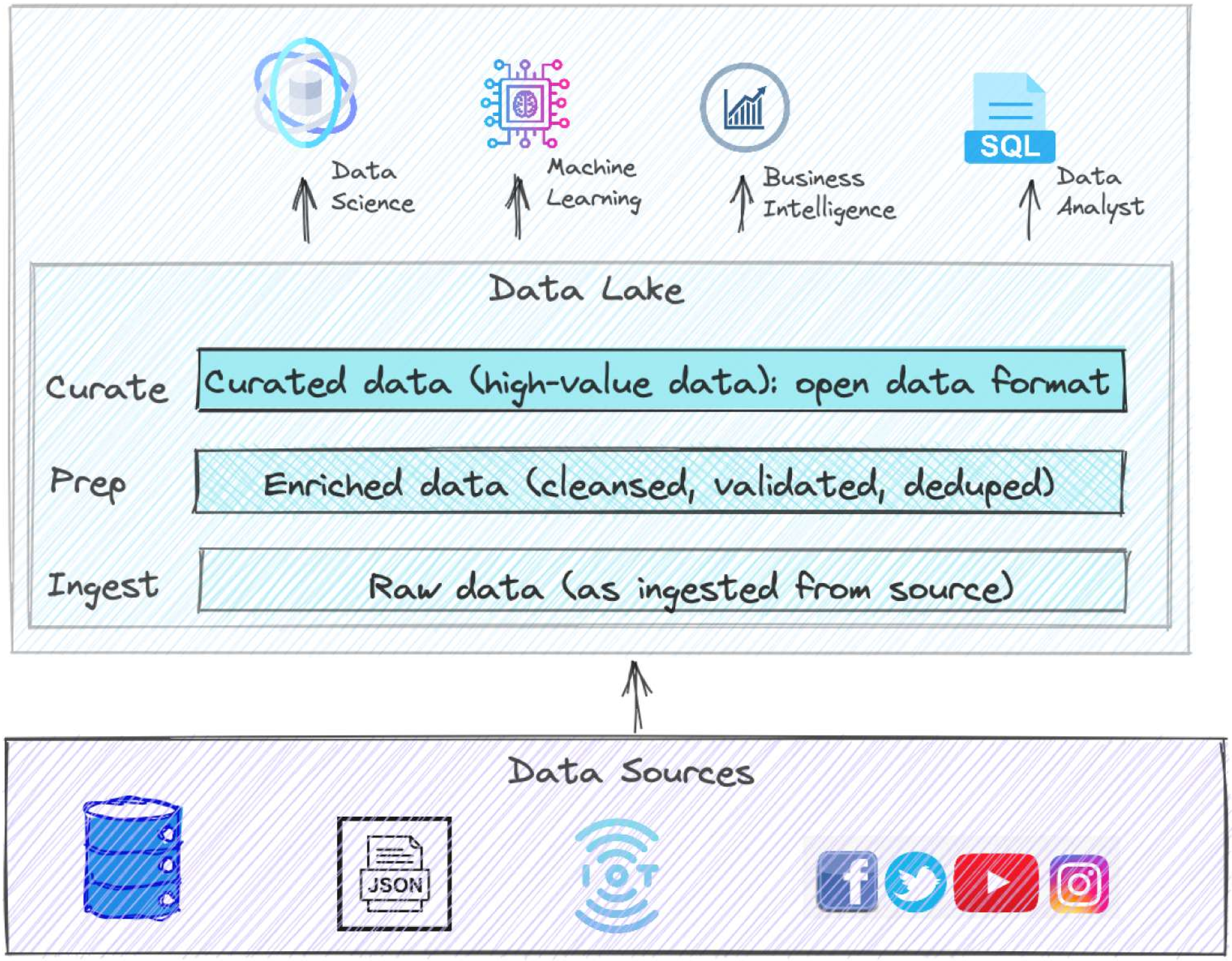
The allure of the lakehouse model stems from

01 Cost-effectiveness due to the affordability of data lakes compared to data warehouses.

02 Eliminating the need to transfer data between the lake and the warehouse.

02 Facilitating seamless data sharing between data scientists and BI teams.

Lakehouse Reference Design offers a basic depiction of the lakehouse design, emphasizing the absence of a separate cloud data warehouse.


Data Science
Machine Learning
Business Intelligence
SQL
Data Analyst
Data Lake
Curate
Curated data (high-value data): open data format
Prep
Enriched data (cleansed, validated, deduped)
Ingest
Raw data (as ingested from source)
Data Sources
JSON
IOT

Historically, data lakes weren't structured to support BI queries. However, advancements have bridged this gap. Data warehouses prioritize structured data for efficient query processing, while data lakes focus on scalability and versatility.

Data warehouses provide:

Schema definition and enforcement.

ACID-compliant transactions ensuring data integrity.

SQL optimization for BI and data analysis tools.

On the other hand, data lakes offer:

- Storage and processing capabilities for unstructured data.
- Cost-effective storage solutions
- Advanced data management features.

Merging these two architectures has been challenging due to their contrasting strengths and weaknesses. However, the increasing adoption of data lakes and the emergence of technologies like Delta Lake, Apache Iceberg, and Apache Hudi have made the lakehouse model feasible.

These technologies share a common goal: defining data stored in the data lake. They achieve this through:

- Open file formats.
- A metadata layer.
- Compute engines compatible with these formats and metadata.

Lakehouse Architecture Benefits and Challenges:

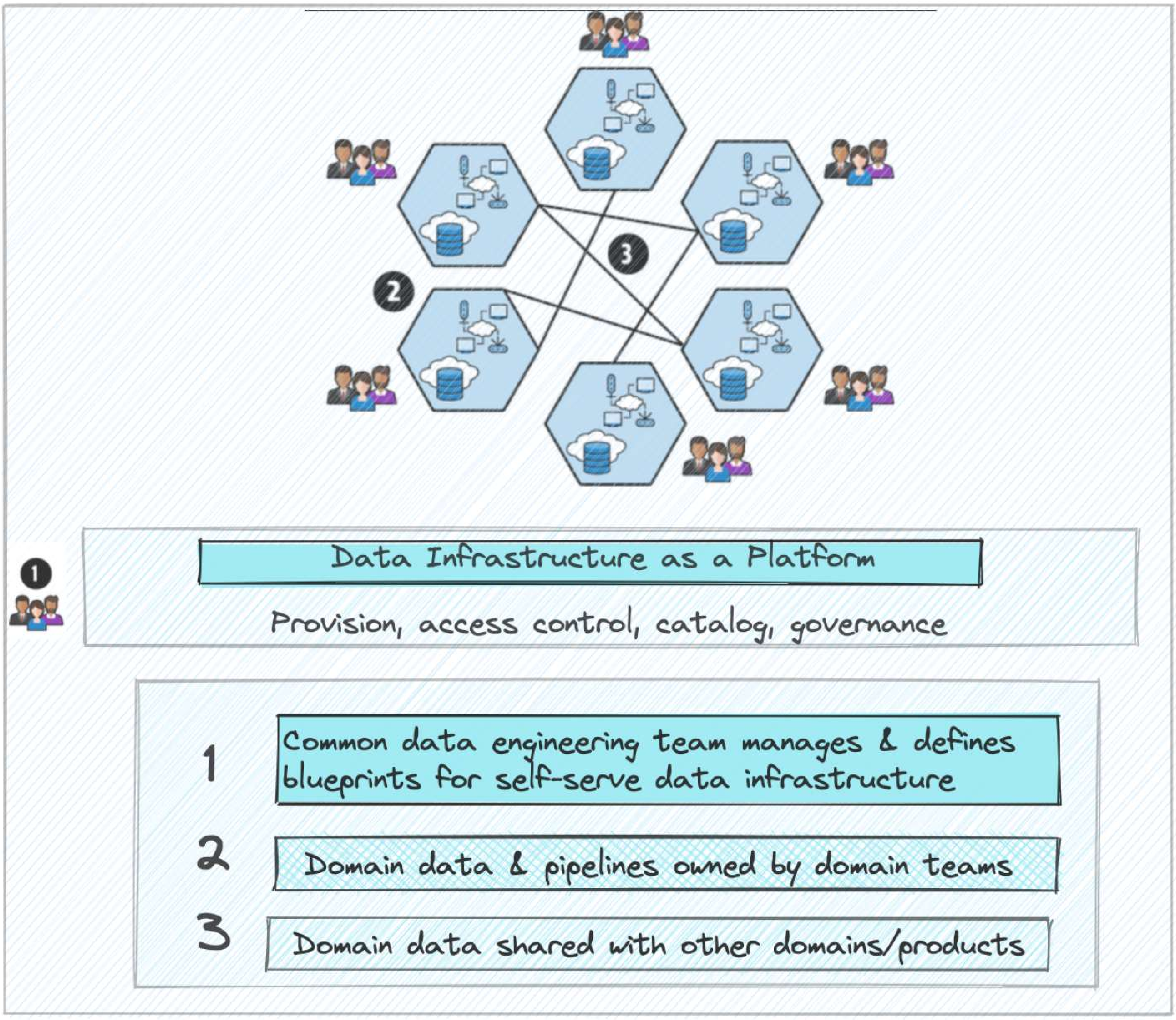
Lakehouses enable efficient BI/SQL operations directly on data lakes, fostering collaboration among different data platform users. This approach is cost-effective compared to traditional data warehouses.

However, building a lakehouse demands meticulous design and architecture. While data warehouses offer an optimized solution, they lack the openness of lakehouses. The rapid evolution in the lakehouse domain suggests that we can expect further simplifications and innovations in the near future.

Data Mesh Architecture

In recent years, the concept of the data mesh has emerged as a revolutionary approach to data architecture and operations. Introduced by Zhamak Dehghani from Thoughtworks in 2019, the data mesh paradigm emphasizes decentralization, treating data as a product rather than just an asset.

Historically, organizations have centralized their data operations, with a singular data platform team overseeing various roles, from data platform architecture and data management to data governance and ingestion. This centralized approach, while efficient in some ways, often leads to bottlenecks, especially as the organization and its data needs grow.



Data Infrastructure as a Platform

Provision, access control, catalog, governance

1 Common data engineering team manages & defines blueprints for self-serve data infrastructure

2 Domain data & pipelines owned by domain teams

3 Domain data shared with other domains/products

Key Principles of Data Mesh

Decentralized Responsibility:
Instead of a central team handling all data-related tasks, responsibilities are distributed across various domain-specific teams.

Distributed Architecture
Moving away from a single, monolithic data storage system to a network of interconnected data storage solutions.

Integrated Data Solutions
Data isn't isolated; it's integrated with relevant business processes and systems.

Federated Governance
Instead of top-down governance, each domain has autonomy while still adhering to overarching organizational policies.

Data as a Product
Data is viewed as a product designed to serve and benefit its users, not just as an asset to be stored.

Advantages of Data Mesh

Scalability:
The decentralized nature of data mesh allows for easy scaling as the organization grows.

Flexibility:
Teams can choose the best infrastructure and technology for their specific needs.

Promotion of Data Culture:
Data becomes an organization-wide focus, not just the purview of a single team.

Potential Challenges

Skill Requirements

The success of the data mesh approach relies on having skilled software developers in each product team.

Complexity

Introducing a distributed layer to data architecture can increase its complexity.

Future Innovations

The data mesh model is still relatively new, and rapid innovations are expected in the coming years to simplify its deployment and management.

Which Architecture to Choose & When

Understanding Your Audience

Every project's success hinges on its alignment with its target audience. Begin by identifying and prioritizing your goals and understanding your primary user groups.

These might include:

01 **BI/Data Analysts:** Aim to curate datasets for them, possibly through scheduled data ingestion and processing tasks.

02 **Data Scientists/Exploratory Analysts:** Establish an infrastructure that allows them to integrate their datasets for analysis. Optionally, you can manage data ingestion from recognized sources.

Align with Business Objectives:

While the allure of new technology is undeniable, it's essential to remember that it's a tool to achieve a broader goal. Some driving factors for organizations considering a cloud data lake include:

01 **Cost Efficiency:** Transitioning to a cloud data lake can significantly reduce total costs. Ensure your architectural decisions align with potential cost savings.

02 **Innovative Applications:** Some organizations are drawn to cloud data lakes to harness emerging technologies like machine learning or real-time analytics. If this resonates with you, ensure your goals reflect this. For instance, are you aiming to enhance user engagement or introduce smarter product features?

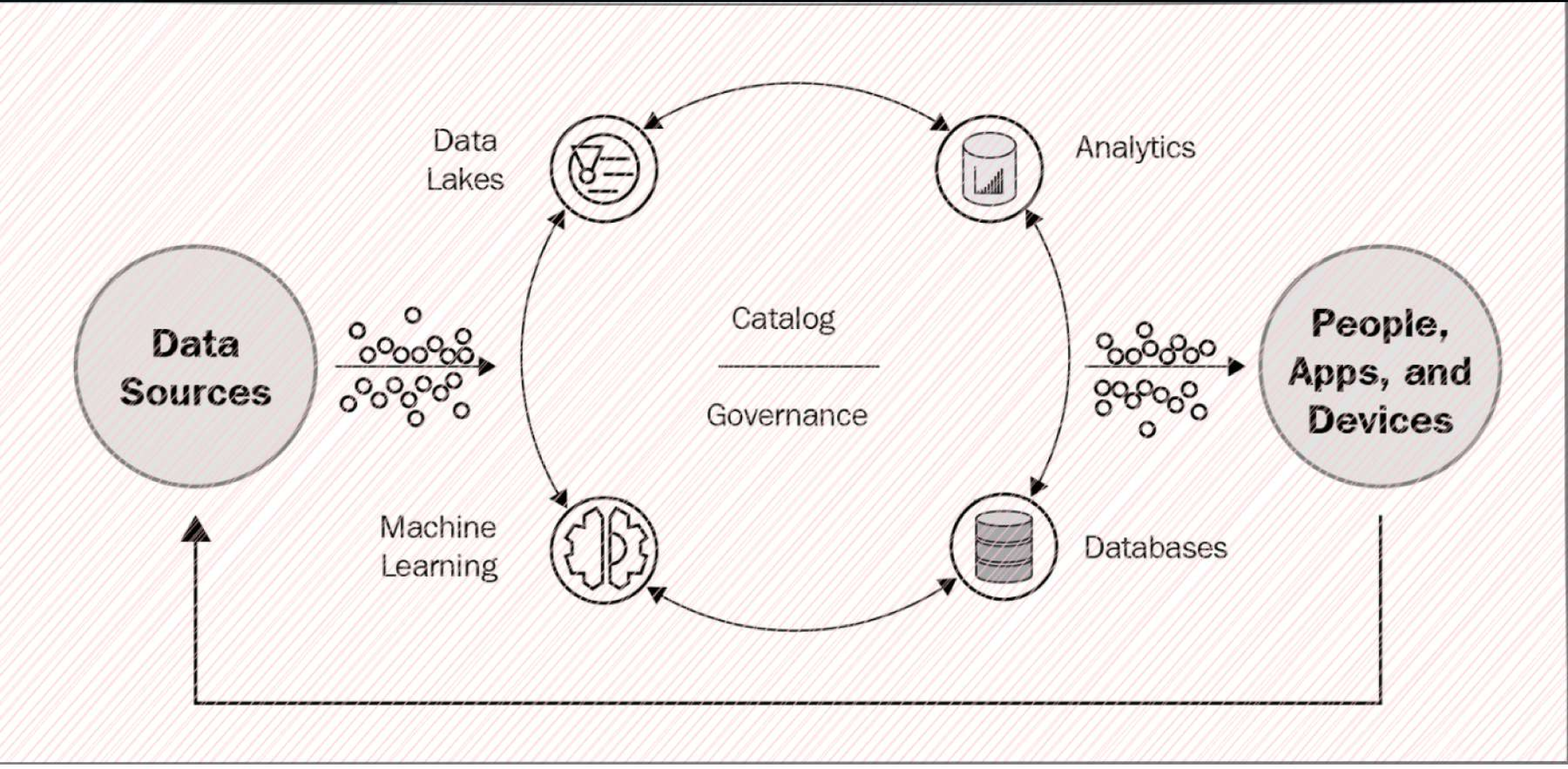
03 **Time Constraints:** Sometimes, external factors, like expiring on-premises hardware support or software licenses, can influence the pace and nature of the transition to the cloud.

Architecture	Cost Assessment	Scenario Adaptability	Development Intricacy	Ecosystem Evolution	Organizational Expertise Needed
Modern Data Warehouse	Moderate, with cost savings from data lake storage, but cloud warehouses remain pricey.	Balanced adaptability with a broader tool ecosystem. Data correlation might require data duplication.	Moderate; requires efficient data lake design with available guidance.	Balanced maturity; high for data warehouse consumption but needs tuning for other scenarios.	Moderate; the team should comprehend organizational needs and design accordingly.
Data Lakehouse	Cost-effective due to unified storage and on-demand compute engines.	Highly adaptable, supporting a wide range of analyses and promoting data sharing.	Ranges from moderate to intricate; selection of datasets and open data format is crucial.	Evolving maturity with technologies like Delta Lake and Apache Iceberg.	Moderate to high expertise; the team should be familiar with emerging technologies.
Data Mesh	Balanced cost; distributed design is economical, but automation and governance need investment.	Highly flexible, supporting diverse architectures without central bottlenecks.	Highly intricate, demanding comprehensive automation and scalable design.	Minimal: relatively nascent in guidance and available tool sets.	High expertise; the data platform and product/domain teams need to be skilled up in data lakes.

Modern Data Architecture on **AWS**

A modern data architecture dissolves the traditional barriers between various data systems, ensuring a fluid integration of data lakes, data warehouses, and specialized data repositories. This framework acknowledges that a universal solution often results in trade-offs within the data analytics platform. This integration isn't limited to just data systems; it also includes consolidated governance and streamlined data transitions.

This advanced data framework emerges as a solution to challenges like rapid data expansion, scalability concerns, governance complexities, data compartmentalization, and escalating costs



Every piece of data an enterprise accumulates is pivotal in reshaping its operations. The swifter an organization can extract insights from this data, the more promptly it can make informed decisions to propel its business trajectory. Yet, as data multiplies in volume and intricacy, it can also decelerate business processes. As data accumulates within data lakes or specialized repositories, transferring this data becomes increasingly challenging, giving data its own inertia.

Thus, within a modern data architecture, there should be systems that facilitate the smooth transition of data, aiming to extract insights using the most suitable tools and services. This data transition can be internal, external, or even shared across various boundaries.

Inside-out

In this pattern, data is initially ingested and refined within the data lake. Subsequently, specific segments of this data, based on the application, are transferred to specialized storage systems. For instance, data from various cloud-based applications is first directed to a data lake. Here, it undergoes ETL processes, and a selected segment is then transferred to a data warehouse for routine analytics.

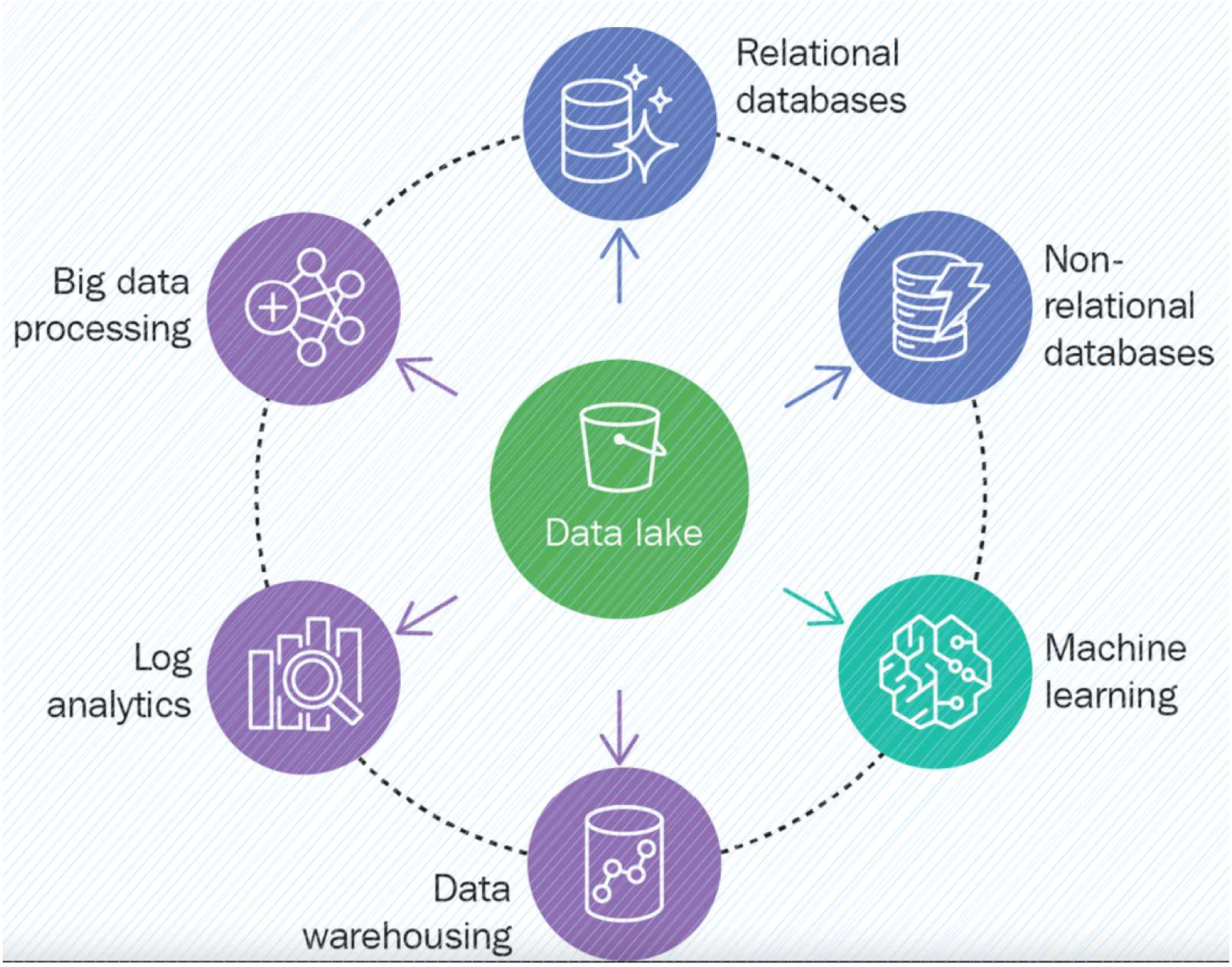


Diagram - Courtesy of AWS

Outside-in

In this pattern, data is first directed into a specialized storage system. Later, it's transferred to a data lake for comprehensive analysis. Taking the three-tier architecture model as an example, application-specific data is stored in a relational database system. This data is later shifted to the data lake for broader analytical purposes. Multiple such specialized systems might contribute data to a central data lake.

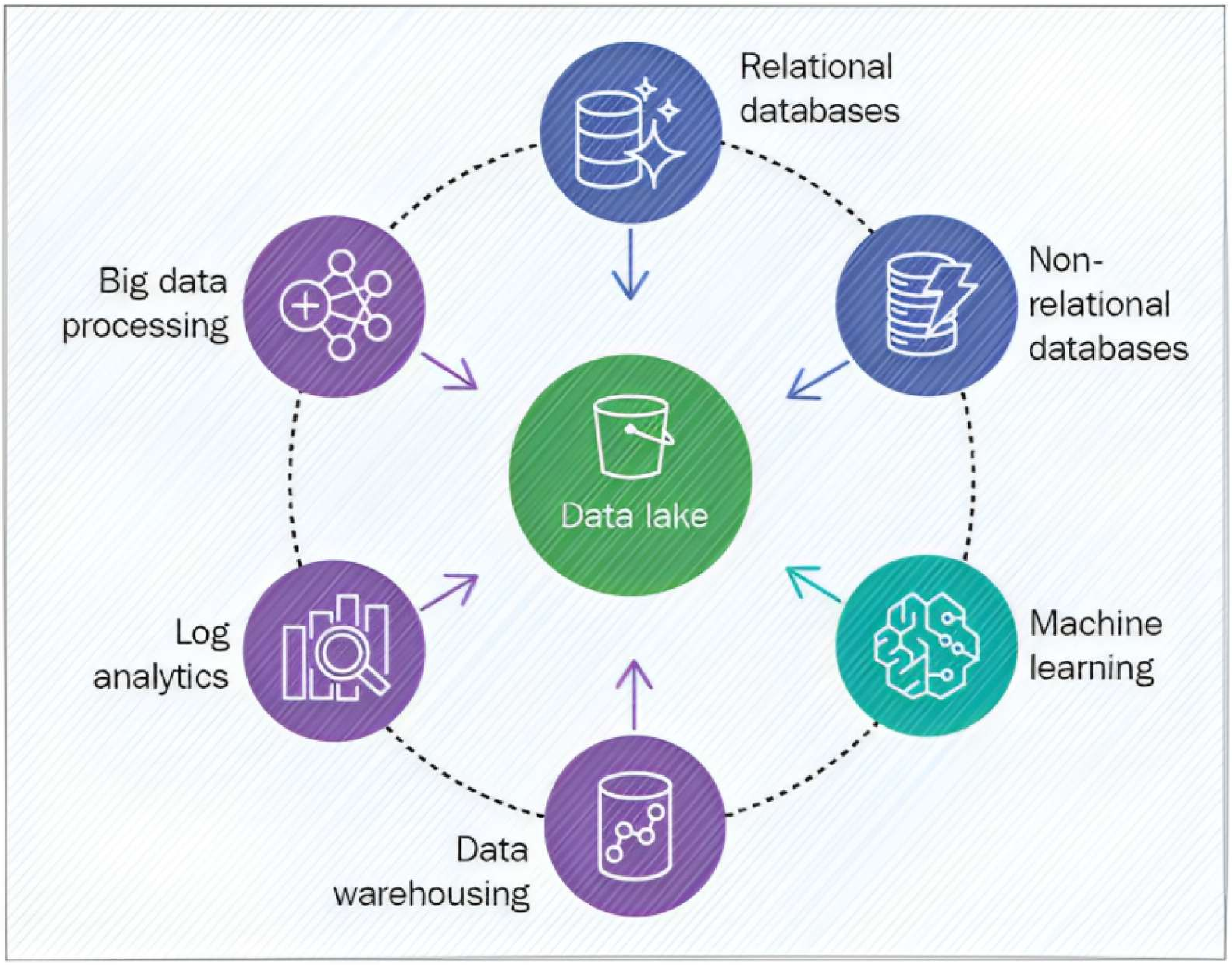


Diagram - Courtesy of AWS

Around the perimeter

In specific scenarios, there's a need to transfer data directly between two specialized storage systems to address particular requirements. As data transitions without necessarily being stored in a data lake, this approach is termed as data movement around the perimeter. For example, in a three-tier architecture, data can directly be loaded from a transactional database into a data warehouse for analytics.

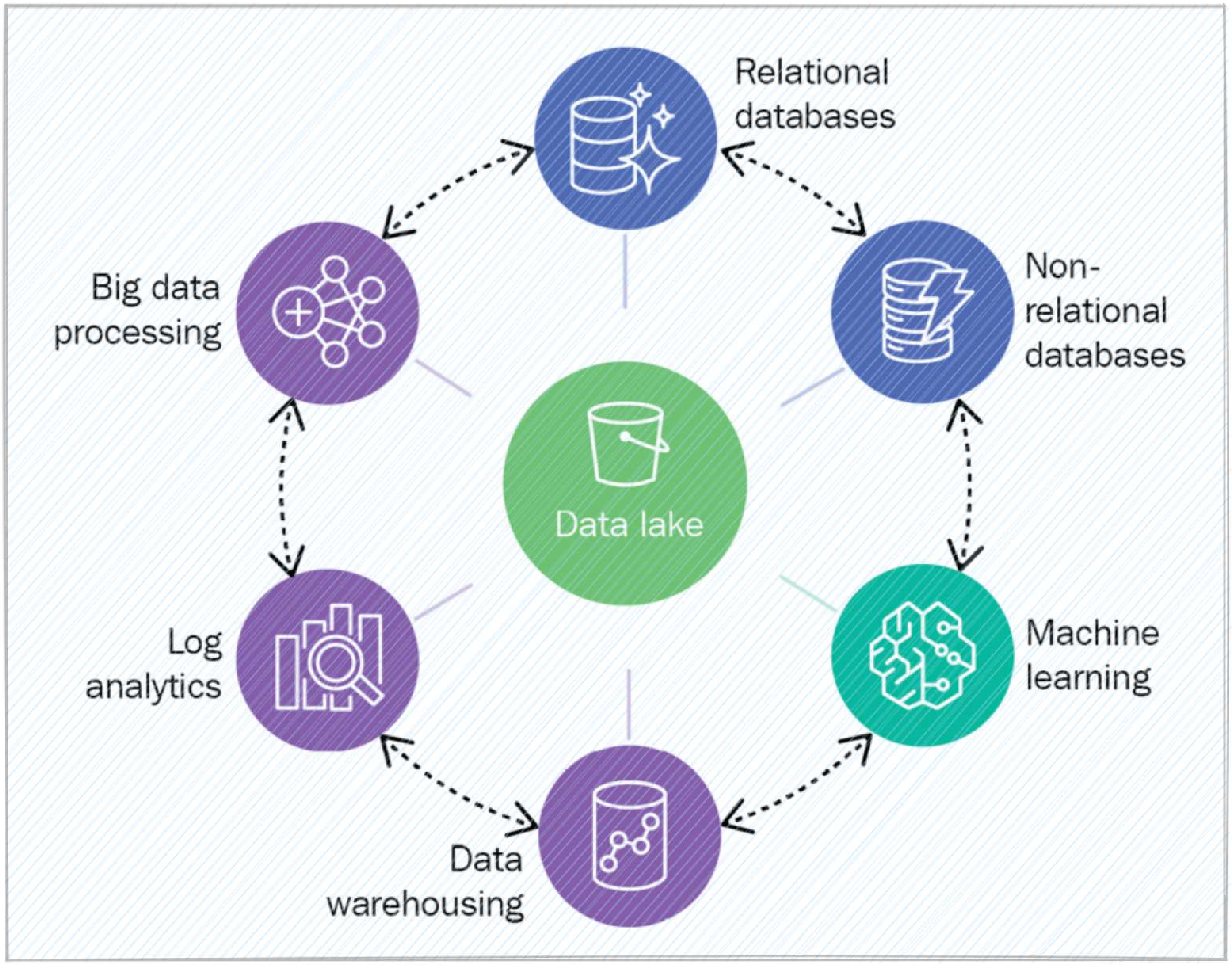


Diagram - Courtesy of AWS

Sharing across

Data's true potential is realized when it's accessible across different sectors. A contemporary data framework promotes the effortless sharing of data, both internally and externally. For instance, data generated in one department might need to be accessible to another, ensuring the entire organization reaps its benefits.

Utilizing AWS's modern data architecture, customers can harness specialized tools and services to construct their data infrastructures. This ensures a consolidated approach to security and data governance. Furthermore, AWS's framework enables organizations to expand their operations cost-effectively without compromising system efficiency. It promotes the seamless sharing of data across departmental boundaries, empowering businesses to make swift, informed decisions at an expansive scale.

To achieve all of this, AWS has provided five pillars for building a modern data architecture. Let's look at them in detail.

Pillars of a modern data architecture

A modern data architecture is required to break down data silos so that data analytics, descriptive as well as predictive using artificial intelligence/machine learning (AI/ML), can be done with all the data aggregated into a central location. In order to meet all the business needs around deriving value out of the data in a fast and cost-effective manner, the architecture requires certain pillars to be in place, as follows:

01 Scalable data lakes

02 Purpose-built analytics services

03 Unified data access, including seamless data movement

04 Unified governance

05 Performance and cost-effectiveness

The following diagram illustrates these pillars for you:

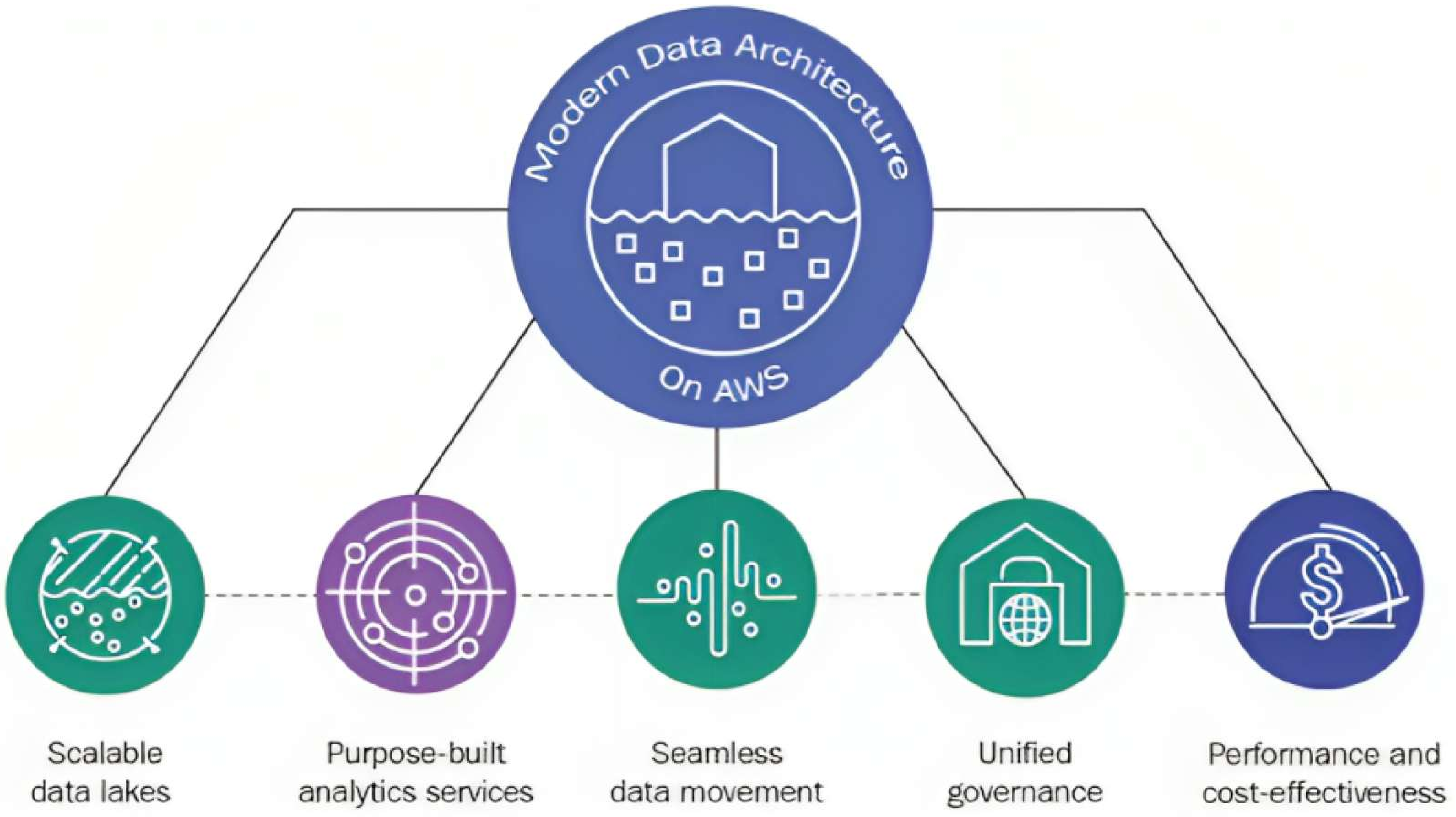


Diagram - Courtesy of AWS

Scalable data lakes

At the heart of any advanced data framework lies the data lake. Given the influx of business data from varied sources, these reservoirs can expand rapidly. Consider the challenges for IT teams in ensuring scalability, reliability, durability, and high availability, all while maintaining performance and cost-efficiency.

Enter Amazon Simple Storage Service (S3). As an object storage solution, S3 offers essential features for efficient data lake management, including scalability, high availability, and swift data access. The subsequent chapter delves deeper into constructing these expandable data reservoirs on AWS, exploring various use cases and design strategies.

The following diagram shows Amazon S3 as the central service for storing all the data in a data lake:

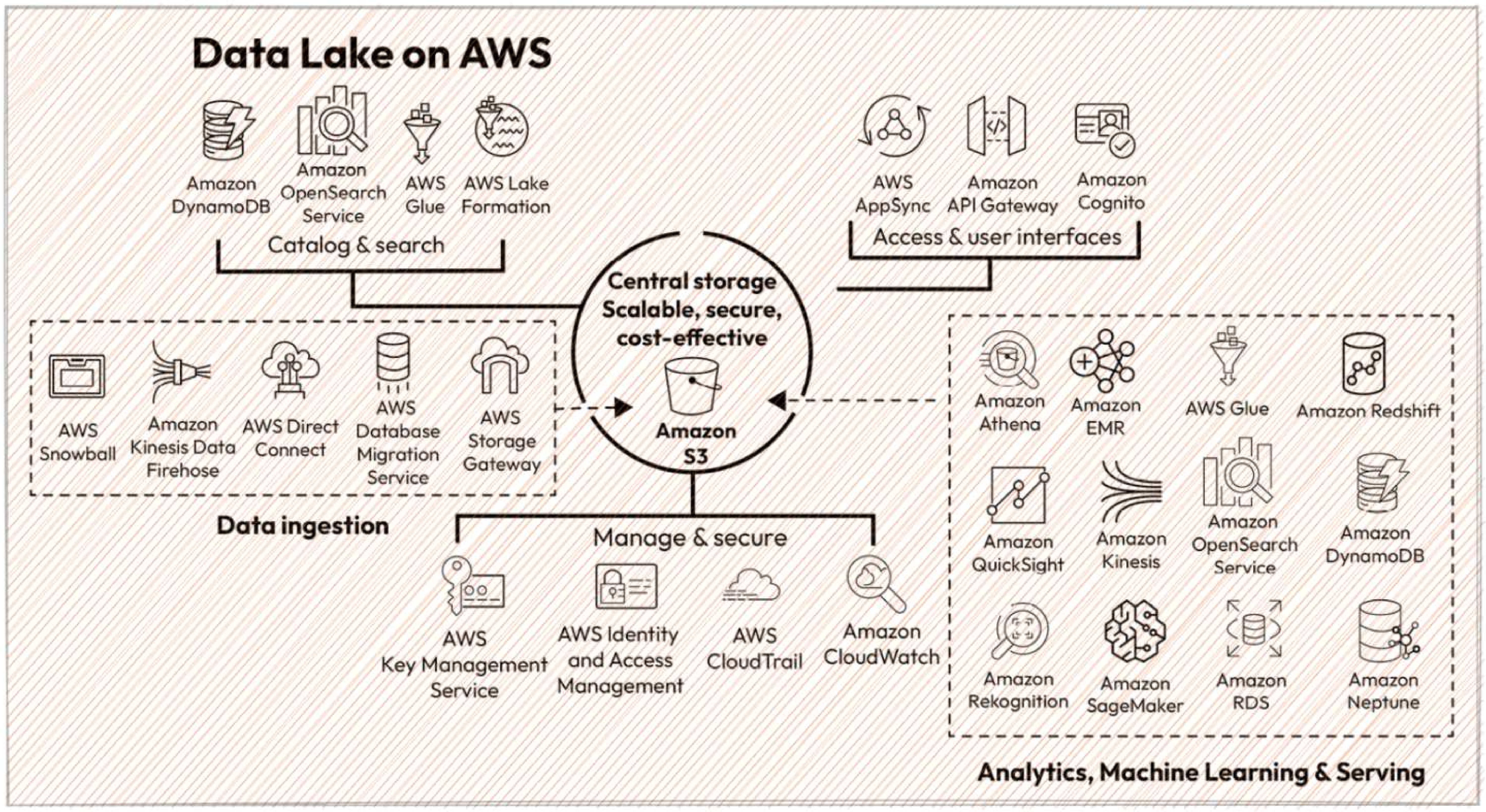
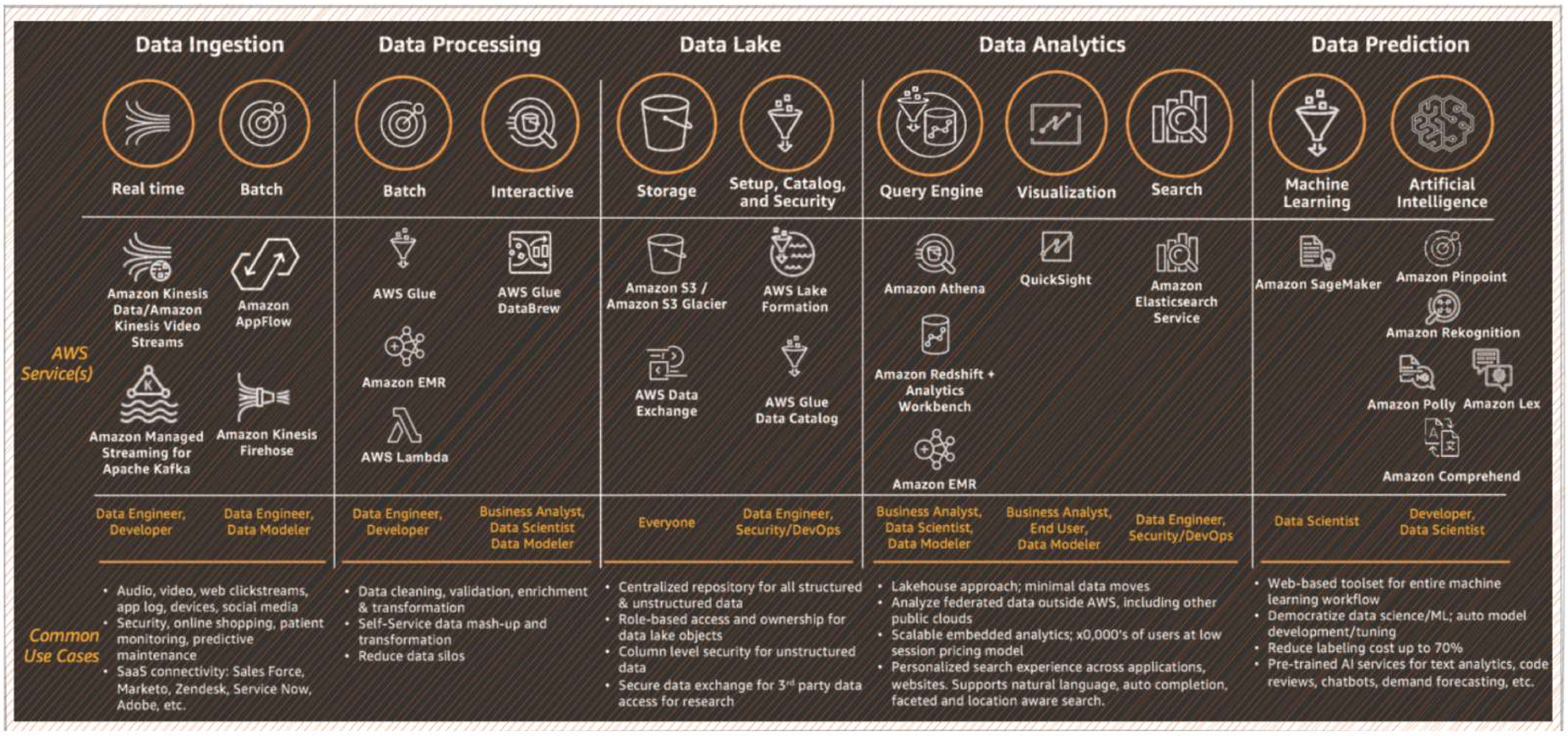


Diagram - Inspired from AWS blog

Purpose-built analytics services

AWS emphasizes the importance of using the most suitable tool for each specific task within its modern data framework. To this end, AWS offers an extensive array of data and analytical services, ensuring optimal cost and performance for distinct use cases.

The following diagram shows the purpose-built services that help to build a modern data platform:



	Data Ingestion		Data Processing		Data Lake		Data Analytics			Data Prediction	
	Real time	Batch	Batch	Interactive	Storage	Setup, Catalog, and Security	Query Engine	Visualization	Search	Machine Learning	Artificial Intelligence
AWS Service(s)	Amazon Kinesis Data/Amazon Kinesis Video Streams Amazon Managed Streaming for Apache Kafka	Amazon AppFlow Amazon Kinesis Firehose	AWS Glue Amazon EMR AWS Lambda	AWS Glue DataBrew	Amazon S3 / Amazon S3 Glacier AWS Data Exchange	AWS Lake Formation AWS Glue Data Catalog	Amazon Athena Amazon Redshift + Analytics Workbench Amazon EMR	QuickSight	Amazon Elasticsearch Service	Amazon SageMaker	Amazon Pinpoint Amazon Rekognition Amazon Polly Amazon Lex Amazon Comprehend
	Data Engineer, Developer	Data Engineer, Data Modeler	Data Engineer, Developer	Business Analyst, Data Scientist Data Modeler	Everyone	Data Engineer, Security/DevOps	Business Analyst, Data Scientist, Data Modeler	Business Analyst, End User, Data Modeler	Data Engineer, Security/DevOps	Data Scientist	Developer, Data Scientist
Common Use Cases	• Audio, video, web clickstreams, app log, devices, social media • Security, online shopping, patient monitoring, predictive maintenance • SaaS connectivity: Sales Force, Marketo, Zendesk, Service Now, Adobe, etc.		• Data cleaning, validation, enrichment & transformation • Self-Service data mash-up and transformation • Reduce data silos		• Centralized repository for all structured & unstructured data • Role-based access and ownership for data lake objects • Column level security for unstructured data • Secure data exchange for 3rd party data access for research		• Lakehouse approach; minimal data moves • Analyze federated data outside AWS, including other public clouds • Scalable embedded analytics; x0,000's of users at low session pricing model • Personalized search experience across applications, websites. Supports natural language, auto completion, faceted and location aware search.			• Web-based toolset for entire machine learning workflow • Democratize data science/ML; auto model development/tuning • Reduce labeling cost up to 70% • Pre-trained AI services for text analytics, code reviews, chatbots, demand forecasting, etc.	

Diagram - Inspired from AWS blog

Unified data access

Just as driving a luxury vehicle offers a unique experience, a contemporary data platform should provide a seamless experience for its users. The processes to populate the platform with data should be straightforward, efficient, and maintainable. Ensuring effortless data transition is paramount, a topic so intricate that it warrants its own whitepaper.

Unified governance

In an ideal world, data security and governance might be redundant topics. However, given the omnipresence of potential threats, the emphasis on consistent oversight and security is paramount.

A modern data architecture on AWS ensures streamlined and flexible data access management. It's crucial to have the capability to track all data access, ensuring adherence to compliance standards. A subsequent chapter will delve into AWS's offerings for robust data governance.

Performant and cost-effective

A data platform's value diminishes if it's either too slow or too expensive. Hence, it's vital to deploy the appropriate features in the right scenarios, ensuring the platform's optimal operation. An upcoming chapter will detail how AWS aids in achieving this balance, highlighting the best practices to ensure a balance between cost and performance.

Conclusion

In the evolving landscape of data management and analytics, the emergence of scalable cloud data lakes and various architectural patterns underscores the importance of adaptability, efficiency, and strategic alignment with business objectives. From the traditional problem-first approach to the revolutionary data mesh paradigm, the journey of data architecture has been marked by innovations aimed at addressing the ever-growing challenges of data volume, complexity, and the need for real-time insights.

Amazon's AWS, with its comprehensive suite of services, stands at the forefront of modern data architecture evolution. It not only facilitates seamless transitions but also guarantees scalability, flexibility, and cost-effectiveness. AWS's Modern Data Architecture emphasizes the integration of data lakes and warehouses, offering a harmonious blend of storage and analytical capabilities. Whether organizations opt for the unified approach of the Data Lakehouse Architecture or the decentralized, product-centric strategy of the Data Mesh, AWS ensures that each architectural pattern is tailored to meet specific organizational requirements, all while maintaining the highest standards of efficiency and performance.

However, the key to successful implementation lies in understanding the target audience, be it BI analysts or data scientists, and ensuring that the chosen architecture aligns seamlessly with the broader business objectives. As organizations continue to navigate the complexities of data management, the focus should remain on harnessing the potential of data as a transformative tool, driving innovation, and ensuring sustainable growth.

References:

https://docs.aws.amazon.com/whitepapers/latest/modern-data-architecture-rationales-on-aws/modern-data-architecture-on-aws.html

https://aws.amazon.com/blogs/architecture/optimize-your-modern-data-architecture-for-sustainability-part-1-data-ingestion-and-data-lake/

https://aws.amazon.com/blogs/architecture/optimize-your-modern-data-architecture-for-sustainability-part-2-unified-data-governance-data-movement-and-purpose-built-analytics/