IKEA Cloud Architecture

YADUMAN SINGH CHOUHAN

April/18/2024

—

Data Science

—

DR. Junaid Qazi

Contents

INTRODUCTION

A sample of cloud architecture for IKEA in Azure. In the digital era, businesses are embracing cloud architecture for its scalability and agility. This report provides a succinct analysis of cloud architecture's components such as Data Factory, Event hub, Pipelines used for ETL & ELT process. We explore the evolution of cloud computing, delineate its layers, and assess the benefits and challenges of cloud adoption. Additionally, we are going to discuss the Data flow in this cloud architecture and get a better understanding of Data storage.

KEY QUESTIONS

1) Flow of the data (source, transformation, destination)?

2) How to handle the flow?

3) Data pipeline (best strategy from creation to monitoring)?

4) How to save cost (typically computer)?

To optimize the flow of data from source to destination, we implement efficient transformations ensuring data integrity and relevance. Handling the flow involves employing robust mechanisms for data validation, error handling, and scalability. Our data pipeline adopts the best strategy from creation, incorporating automated workflows, scheduling, and monitoring for seamless operation. To save costs, we leverage cloud services for flexible scalability, resource allocation optimization, and pay-as-you-go models, thus minimizing expenditure on computing resources.

Components Of Pipeline

DATA SOURCE:

Data analytics relies on diverse sources of data, each offering unique insights and challenges. These sources include:

Databases and Data Warehouses: Traditional repositories structured for storing and managing large volumes of structured data, often used for historical analysis and reporting.
Files: Data stored in various file formats like CSV, Excel, JSON, or XML, offering flexibility in data exchange and integration.
Application Programming Interfaces (APIs): Interfaces provided by software applications, allowing access to their data and functionalities programmatically, facilitating real-time data retrieval and integration.
Sensor Data from Internet of Things (IoT) Devices: Streams of data generated by sensors embedded in physical objects or environments, offering real-time insights into operations, environment, and user behavior.

Social Networking and Other Software-as-a-Service (SaaS) Sites: Platforms hosting vast amounts of user-generated content, interactions, and transactions, providing valuable insights into customer behavior, market trends, and social dynamics.
These diverse data sources play a crucial role in fueling analytics-driven decision-making processes across industries and domains.

DATA INGESTION:

Data ingestion is the vital process of collecting, transferring, and consolidating raw data into a usable format for analysis and storage. Azure Data Factory, a robust cloud-based service, streamlines this critical task by orchestrating and automating data workflows across various sources and destinations. Through its intuitive interface and powerful integration capabilities, Data Factory efficiently extracts data from disparate sources such as databases, applications, and files, transforming it on-the-fly to fit desired schemas or structures. Leveraging Azure Data Factory's scalability, organizations can seamlessly ingest vast volumes of data, ensuring timely access to valuable insights for decision-making. By facilitating seamless connectivity and data movement, Data Factory empowers businesses to harness the full potential of their data ecosystem, driving innovation and competitiveness in the digital age.

DATA TRANSFORMATION:

Data transformation stands as a pivotal stage in data processing, its intricacy varying greatly depending on diverse factors. The complexity emerges from a confluence of variables, notably the multiplicity and diversity of data sources involved. Furthermore, the level of sophistication required in implementing business logic for tasks like data cleaning, declassification, and aggregation significantly contributes to the complexity spectrum. Another determinant is the frequency of data ingestion and subsequent transformations, as real-time or batch processing necessitates distinct approaches. Moreover, the scale of ingested data, along with the temporal extent of aggregation—whether spanning days, weeks, months, or even years—adds layers of intricacy to the transformation pipeline. In navigating these complexities, organizations must employ robust tools and methodologies to ensure the accuracy, efficiency, and reliability of their data transformation processes, ultimately underpinning informed decision-making and strategic initiatives.

DATA DESTINATION:

A crucial aspect lies in defining data destinations, which serve as repositories for the outcomes of data transformations. These destinations are categorized into two primary types: final and intermediate. The destination encapsulates the ultimate results of transformations, akin to the transformation layer in traditional architectures or the gold layer in the innovative Lakehouse paradigm. It serves as the definitive storehouse of processed data, ready for consumption by end-users or downstream applications. Conversely, intermediate destinations house interim results of transformations, playing a pivotal role in data processing pipelines. In traditional setups, these are referred to as the raw or curation layer, whereas in the emerging Lakehouse model, they are denoted as the bronze or silver layer. These intermediate destinations serve as staging grounds for refining and preparing data before it reaches its final, consumable form, ensuring accuracy, reliability, and accessibility throughout the data lifecycle.

Different Zones Of DATA

BRONZE LAYER:

In cloud architecture, particularly within the context of data management frameworks like the Lakehouse architecture, the bronze layer serves as an intermediate stage in the data processing pipeline. This layer is responsible for storing raw or minimally processed data ingested from various sources before undergoing further refinement and transformation. The term "bronze" reflects the raw and unpolished nature of the data at this stage, analogous to the raw material waiting to be shaped and refined into a valuable product.

In practical terms, the bronze layer typically encompasses data lakes or storage repositories where raw data is deposited without significant alteration or structuring. This layer allows organizations to preserve the fidelity of their original data while providing a foundation for subsequent processing steps. By maintaining the integrity of the raw data, the bronze layer facilitates data lineage and ensures transparency in the transformation process.

SILVER LAYER:

In cloud architecture, particularly within the context of data management frameworks like the Lakehouse architecture, the silver layer represents an intermediate stage between raw data ingestion and final processed data storage. This layer is where data undergoes initial refinement, cleansing, and structuring before it is deemed ready for consumption by downstream applications or analytical processes.

The term "silver" denotes a level of refinement beyond the raw data found in the bronze layer but still retains flexibility for further processing. Data in the silver layer is often transformed into a more standardized format, with basic cleansing operations applied to address issues such as missing values, duplicates, or inconsistencies. Additionally, data in the silver layer may be organized into logical schemas or data models that facilitate easier querying and analysis.

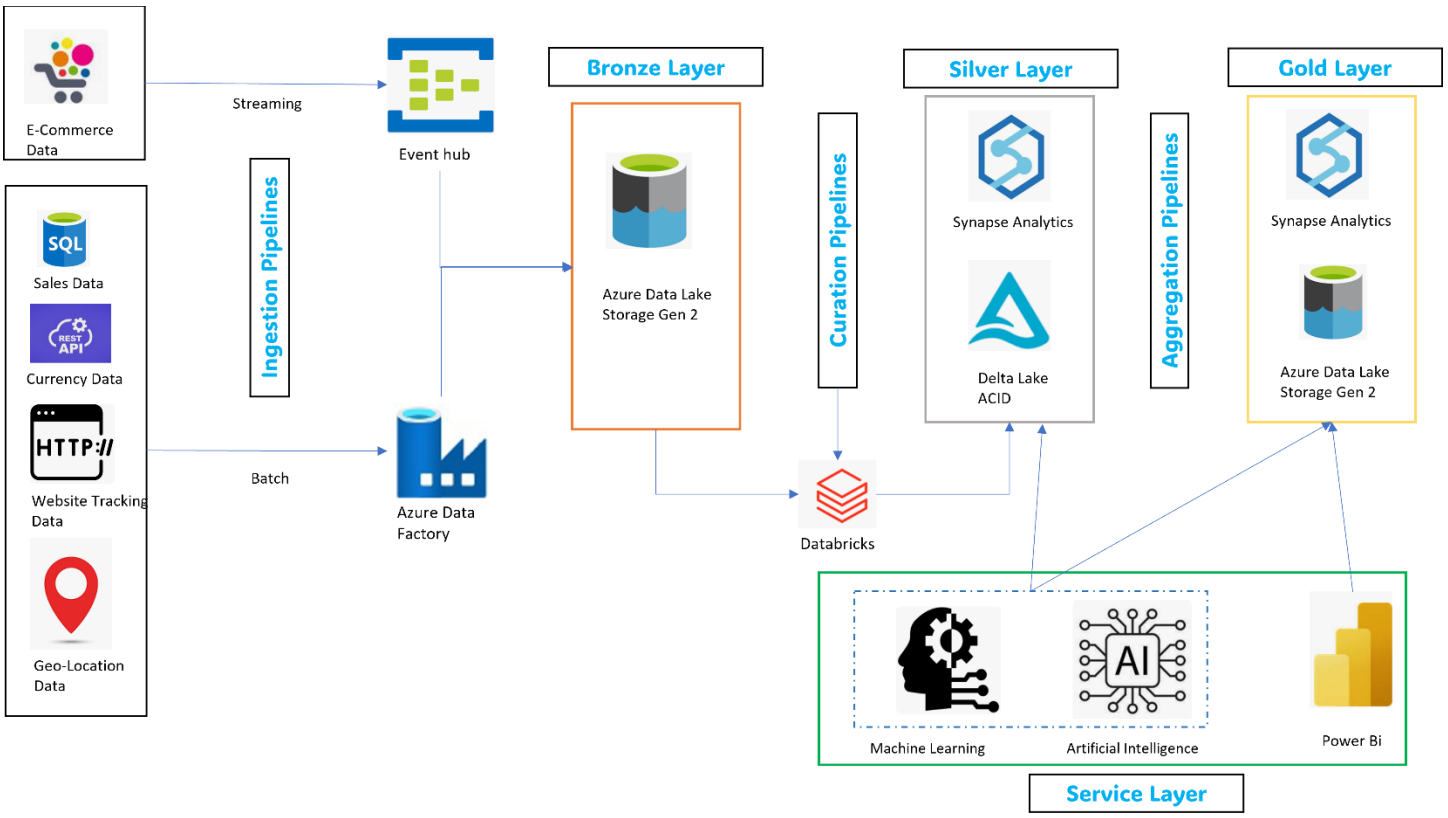
Practically, the silver layer may comprise storage solutions like data lakes or databases equipped with tools for data transformation and preparation. By residing in the silver layer, data becomes more accessible and amenable to exploration and analysis by data scientists, analysts, and other stakeholders.

GOLD LAYER:

In cloud architecture, particularly within the context of data management frameworks like the Lakehouse architecture, the gold layer represents the final stage in the data processing pipeline where refined and processed data is stored for consumption by end-users, applications, or analytical tools. The term "gold" signifies that the data stored in this layer has undergone comprehensive refinement, transformation, and enrichment, resulting in high-quality, valuable insights ready for use in decision-making processes and business operations. Data in the gold layer is typically structured, aggregated, and optimized for specific use cases or analytical requirements. Practically, the gold layer often comprises data warehouses, data marts, or specialized databases optimized for analytics and reporting purposes. These storage solutions are designed to provide fast query performance and support complex analytical operations, enabling users to derive meaningful insights from the data with minimal latency. By residing in the gold layer, data becomes a trusted source of information, serving as the foundation for strategic decision-making, operational optimization, and innovation within organizations. Additionally, the gold layer facilitates data governance, security, and compliance efforts by ensuring the integrity, accuracy, and reliability of stored data assets.

SERVICE LAYER:

In cloud architecture, the service layer typically refers to the component responsible for providing various services and functionalities to support the overall operation of the cloud environment. This layer encompasses a wide range of services offered by the cloud provider, which may include computing resources, storage solutions, networking capabilities, security features, and management tools. The service layer plays a crucial role in enabling users to deploy, manage, and scale their applications and infrastructure efficiently within the cloud environment. It abstracts the underlying hardware and infrastructure complexities, allowing users to focus on developing and running their applications without worrying about the operational overhead associated with managing physical resources.
Overall, the service layer in cloud architecture serves as the foundation for building and operating cloud-based applications and services, providing users with the necessary tools and infrastructure to meet their business requirements efficiently and cost-effectively.



Conclusion

In conclusion, the cloud architecture devised for IKEA features a straightforward data flow. It initiates at the data sources, collecting information from diverse origins, then directing it to Azure Data Factory and the bronze layer via the ingestion pipeline. Here, the data remains stored temporarily in its raw state. Subsequently, the data proceeds to the curation zone via curation pipelines, where it undergoes analysis and transformation. It then advances to the gold layer for further refinement before entering the service layer, facilitating easy display through dashboards to various departments. Notably, the machine learning and artificial intelligence department also has access to the silver zone for additional insights.