Decoding the concept of

DATALAKEHOUSE VS DATAWAREHOUSE

Vikranth Ale

Introduction

Problem Statement

- Many companies rely on data to drive critical business decisions and to serve customers better.
- But the way data is collected, processed, transformed and stored is always challenging as the amount of data captured by the companies is higher than ever before.
- So, there is a need to properly channel data from diversified sources, manage and store it in a robust way for various business needs.

Introduction

Solution

Data Lake and Data Warehouse are the most widely used data storage and management architectures for big data problems in most enterprises today.

Both these architectures are used across industries, and they have their own advantages and disadvantages.

Overcoming the challenges and combining the best of both worlds, researchers have paved way for a new data storage architecture called Data Lakehouse.

One-on-One: Data Warehouse Vs Data Lakehouse

- It is a centralized and structured storage system
- Works well with structured data
- Optimal for data analytics and business intelligence use cases
- Stores data collected from various source in a unified schema
- Storage is costly and time consuming
- Can be Rigid and challenging to scale
- Require upfront data modelling and schema design.

- Combines the best of of Data lake and Data warehouse architecture
- Works well with structured, semi-structured and unstructured data
- Suitable for both data analytics and machine learning workloads
- Stores data in a native format
- Storage is cost-effective, fast, and flexible
- Can scale, store diverse and large-scale data
- No upfront schema design is required

Data Lakehouse : Benefits

Reduced data redundancy: Reduces the data duplication by providing a single all-purpose data storage platform

Cost-effectiveness: Provides cost-effective storage features of data lakes by utilizing low-cost object storage options

Support for a wider variety of workloads: Provide direct access to business intelligence tools like Tableau, PowerBI etc. to enable advanced analytics.

Ease of data versioning, governance, and security: Enforces schema and data integrity to implement robust data security and governance mechanisms.

Data Lakehouse : Main Enablers

The main enablers of Data Lakehouse are the modern cloud-based data platforms and technologies like AWS, Azure, GCP etc

These cloud platforms provide:

- Scalable and cost-effective storage and computing resources
- Secure and durable object storage for storing raw data
- Automated data cataloguing and schema discovery
- Enable serverless and on-demand query processing

Data Lakehouse : Reference Architecture

We can leverage one of the available cloud-providers like to build a highly scalable and highly available Data Lakehouse architecture to meet the growing demands of businesses.

Some services that can be used based on requirements are:

Amazon S3: Store raw data in its native format in an S3 bucket, creating a data lake for all data sources.

AWS Glue Data Catalog: Use AWS Glue to automatically catalog and discover data in the S3 bucket, providing a metadata repository.

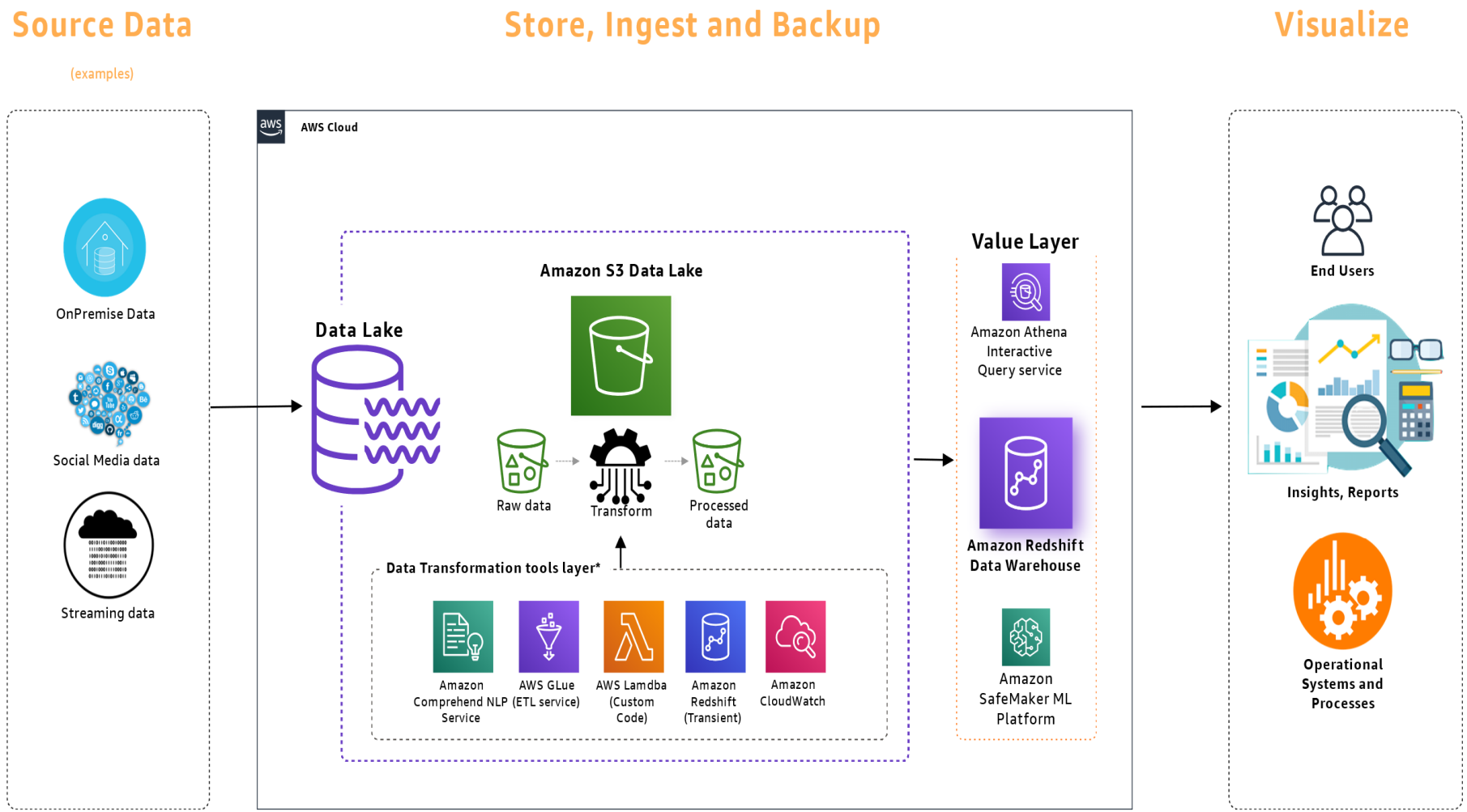
AWS Glue ETL Jobs: Create ETL (Extract, Transform, Load) jobs using AWS Glue to transform raw data into structured and optimized formats as needed.

AWS Lake Formation: Use Lake Formation to define access controls and permissions for different users accessing the data lake.

Amazon Redshift: For optimized and scalable analytics, use Amazon Redshift for data warehousing and structured query processing.

AWS Athena or Amazon Redshift Spectrum: For data exploration and ad-hoc queries, use Athena or Redshift Spectrum to perform serverless and on-demand querying on data in S3 without requiring data movement.

Data Lakehouse : Reference Architecture



Source: https://aws.amazon.com/blogs/architecture/benefits-of-modernizing-on-premise-analytics-with-an-aws-lake-house/

Thank you !