**Project Objectives:**

1. Data Ingestion: Develop a robust mechanism for acquiring data from diverse sources.

2. ETL System: Transform raw data into the appropriate format for seamless processing.

3. Data Lake: Establish a centralized repository to accommodate data from multiple sources.

4. Scalability: Ensure the scalability of our system to handle increasing data volumes effectively.

5. Cloud Integration: Utilize cloud infrastructure, specifically AWS, to process vast amounts of data beyond the capabilities of local computing resources.

6. Reporting: Construct a comprehensive dashboard to extract insights and answers to predefined questions.

**Selected AWS Services:**

1. Amazon S3: Leverage this object storage service for manufacturing scalability, data availability, security, and performance.

2. AWS IAM: Implement identity and access management to securely manage access to AWS services and resources.

3. QuickSight: Utilize Amazon QuickSight, a scalable, serverless, embeddable, machine learning-powered business intelligence service designed for the cloud.

4. AWS Glue: Employ this serverless data integration service to facilitate the discovery, preparation, and consolidation of data for analytics, machine learning, and application development.

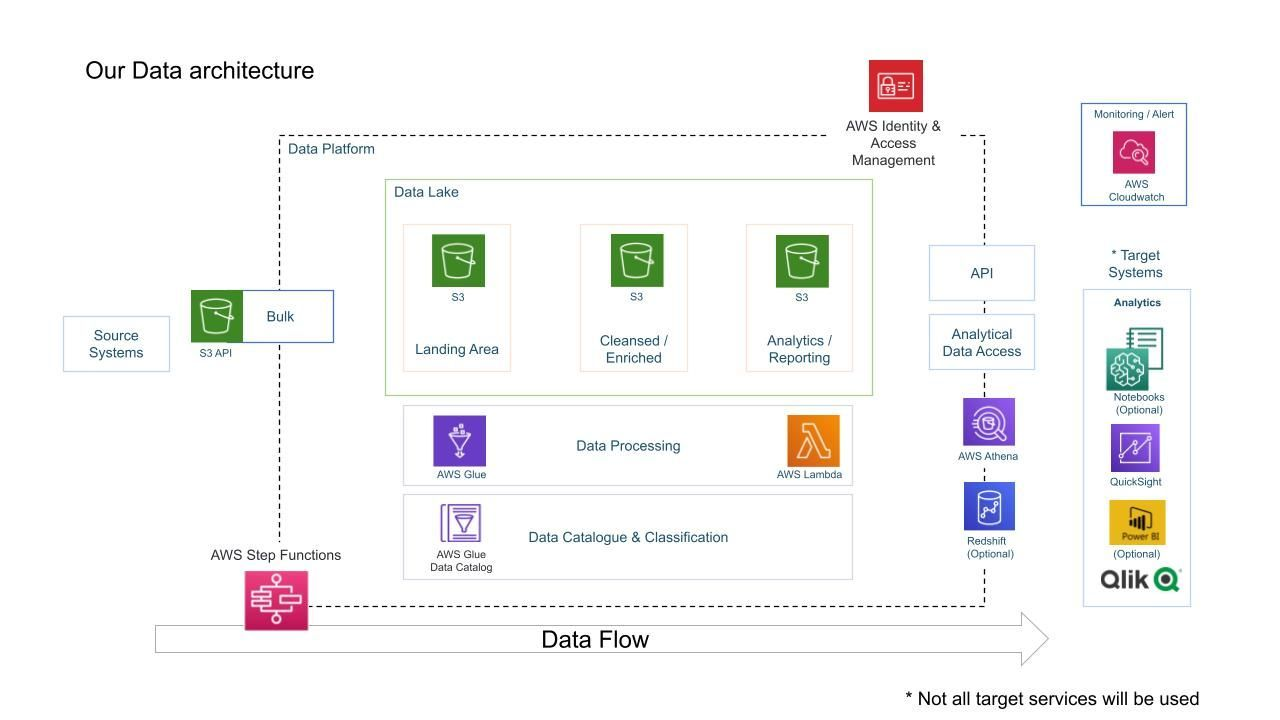
5. AWS Lambda: Leverage Lambda, a computing service allowing code execution without server management.

6. AWS Athena: Implement Athena, an interactive query service for S3, eliminating the need to load data as it remains in S3.

Dataset Description:

The Kaggle dataset under consideration comprises daily statistics (CSV files) on popular YouTube videos spanning several months. Each day witnesses the publication of up to 200 trending videos across various locations. Data for each region is stored in separate files, encompassing video title, channel title, publication time, tags, views, likes and dislikes, description, and comment count. Additionally, a region-specific category\_id field is included in the JSON file linked to each area.

**https://www.kaggle.com/datasets/datasnaek/youtube-new**

****