**PROJECT PROPOSAL**

Decoding YouTube: A Comprehensive Data Analysis

**Team Members**

Yashkumar Subhashbhai Patel Joshi Jay Vijaybhai

(200483929) (200485155)

**Problem Statement**

To conduct a comprehensive data science analysis of YouTube to uncover key patterns, trends, and correlations within the data. This analysis involves exploring various aspects, such as user engagement metrics, video metadata, content categorization, audience segmentation, and viewership patterns. By addressing this problem, we aim to gain a deeper understanding of YouTube's ecosystem, identify factors that contribute to the success of videos and channels, and provide actionable insights to content creators, marketers, and platform administrators. The data generated through YouTube has a huge volume, where using capabilities of cloud computing can make it much easier and efficient.

**Project Goals and Estimated Achievements**

The main objective of this project is to create a reliable and effective system that makes it possible to manage structured and semi-structured YouTube video data effectively and to process it quickly while also allowing for meaningful analysis. The system's primary tasks will be classifying videos according to their content and collecting relevant trending indicators. By reaching this objective, the project hopes to offer users a solid platform for safely handling and analyzing YouTube videos, eventually supporting better decision-making, an improved user experience, and insightful information for different stakeholders.

implementations of reporting tools and visualisations that portray the analysed data in a clear and understandable way. With the help of this visualisation, stakeholders may quickly understand the findings and take appropriate action in light of the knowledge gleaned from the YouTube video data. The ultimate goal is to create a system that can employ cloud computing (AWS) to manage massive amounts of data over time.

**Approach**

We divided this project in to six parts, which are mentioned the below :

1-Data Ingestion: The process of collecting, importing, and obtaining data from numerous sources into a data repository or storage system. Data from many sources, such databases, files, streaming platforms, APIs, sensors, or other data streams, must be collected and brought to a central place for additional processing, analysis, and storage.

2-ETL system: The goal of an ETL system is to extract data from diverse sources, convert it into a standardized and useable format, and load it into a destination data repository, such as a data warehouse or a data lake.

3-Data Lake: A data lake is a centralized storage repository that enables businesses to store and manage enormous volumes of raw or natively formatted organized, semi-structured, and unstructured data. It offers a scalable and affordable method for handling and storing massive amounts of data without the requirement for data transformation or prior schema creation.

4-Scalability: The capacity of a system, method, or process to manage growing data quantities, computing needs, and user expectations without sacrificing performance, efficiency, or accuracy is referred to as scalability. It entails developing and putting into practice solutions that can handle expanding datasets, rising computing complexity, and expanding user bases.

5-Cloud: Difficulty to process large amount of data so in order to process huge volumes different cloud services are used.

6-Reporting: Creation of a Dashboard with the help of visualization techniques which helps the users to answers their questions.

By following this approach, we can effectively collect, process, and analyze YouTube data for our project while ensuring scalability and leveraging cloud services for handling large data volumes.

**Architecture**

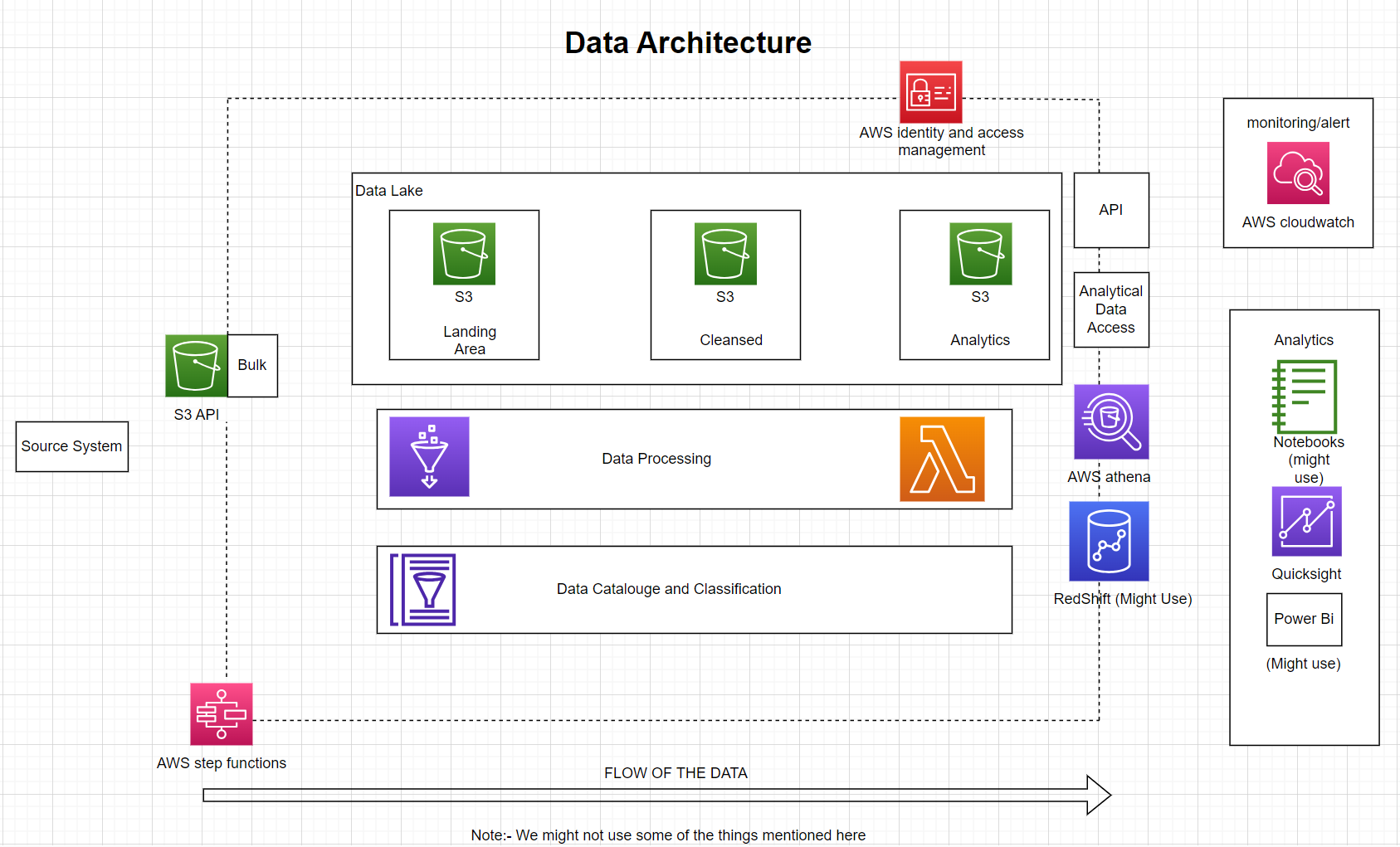


Fig. 1.0

**Literature Review**

* Khanam, Sana, Safdar Tanweer, and Syed Sibtain Khalid. "Youtube Trending Videos: Boosting Machine Learning Results Using Exploratory Data Analysis." The Computer Journal 66.1 (2023): 35-46.
* Eliganti, Ramalakshmi, and A. Reddy. "YouTube Data Analysis & Prediction of Views and Categories." Sharvani, YouTube Data Analysis & Prediction of Views and Categories (April 6, 2022) (2022).
* Khan, M. Laeeq, and Aqdas Malik. "Researching YouTube: Methods, tools, and analytics." The sage handbook of social media research methods (2022): 651-663.

**The Things that we'll be utilising**

* AWS Glue:-

an easy-to-use server less data integration solution that facilitates the discovery, preparation, and combination of data for analytics, machine learning, and application development.

* Amazon S3:-

Manufacturing scalability, data availability, security, and speed are all features of the object storage service Amazon S3.

* AWS IAM:-

We can safely control access to AWS services and resources thanks to identity and access management.

* AWS Athena:-

There is no need to load data into S3 while using Athena, an interactive query service for S3.

* QuickSight:-

A scalable, embeddable, serverless, machine learning-powered business intelligence (BI) service designed for the cloud is Amazon QuickSight.

* AWS Lambda:-

Programmers may run code using the computing service Lambda without building or maintaining servers.

**Schedule**

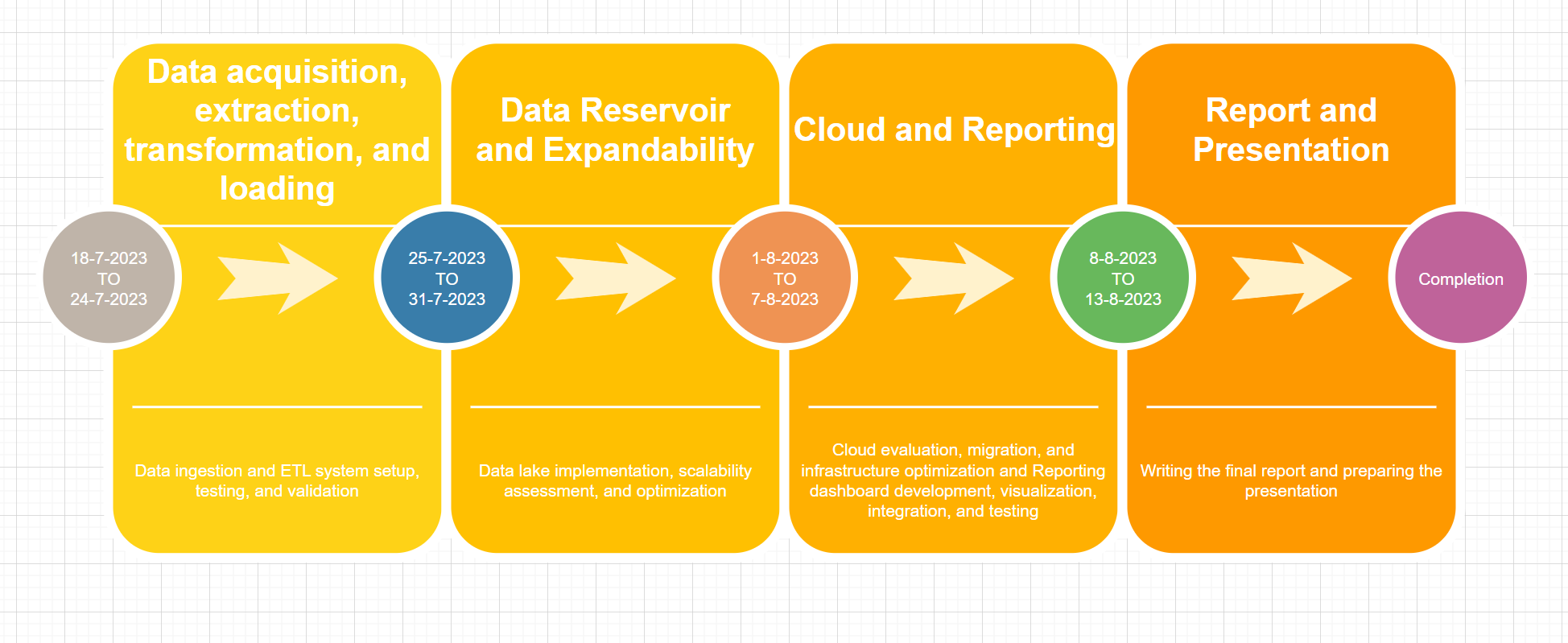


Fig.2.0

**References**

* To understand the information regarding AWS:-

<https://www.youtube.com/watch?v=IT1X42D1KeA>

* For big data related topics:-

<https://www.youtube.com/watch?v=9QxZhapbo0o>

* Literature Review-

<https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=youtube+videos+data+analytics&oq=youtube+videos+data+analytu>

<https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&as_ylo=2022&q=data+analysis+of+Youtube&btnG=>

<https://scholar.google.com/scholar?hl=en&as_sdt=0,5&qsp=2&q=analysis+of+youtube+videos&qst=br>