YouTube Trending Videos ETL Project Report

Prepared by: AMAAN ALI
Date: 10-02-2025

---

1. Introduction

This report outlines the implementation of an **ETL (Extract, Transform, Load) pipeline** for retrieving and processing trending videos from YouTube. The goal of this project is to collect trending video data, transform it into a structured format, and save it for further analysis.

The YouTube API was utilized to fetch trending video data, which was then processed and stored in a tab-separated values (TSV) file. This report provides an overview of the methodology, findings, and future improvements.

---

2. Data Extraction

The **YouTube Data API v3** was used to fetch trending video data. The API request parameters included:

- `part`: `snippet, statistics, contentDetails` (to extract video details, statistics, and duration)
- `chart`: `mostPopular` (fetches trending videos)
- `regionCode`: User-specified country code (e.g., `IN` for India)
- `maxResults`: Number of videos to retrieve (default: 10)
- `key`: API Key for authentication

The extraction function made an HTTP request to YouTube's API endpoint and retrieved trending video data in JSON format.

Challenges:

- API quota limitations restrict the number of requests per day.
- Some videos may have restricted statistics (e.g., missing likes or comments).

---

3. Data Transformation

The raw JSON response was processed into a structured Pandas DataFrame. Key transformations included:

- **Extracting essential video details**: Title, Channel Name, Published Date, Views, Likes, Comments.

- **Parsing ISO 8601 Duration**: Converted video duration from `PT#H#M#S` format into `HH:MM:SS`.
- **Calculating Time Since Published**: Derived the number of **days and hours** since the video was published.
- **Handling Missing Data**: Ensured missing values (e.g., likes, comments) were set to `0` if unavailable.

Example Output Format:

Video ID	Title	Channel	Published At	Days Since Published	Hours Since Published	Views	Likes	Comments	Video Duration
XYZ123	Trending Video 1	Channel A	2025-02-01T10:00:00Z	5	12	1,000,000	50,000	5,000	00:10:45
ABC789	Trending Video 2	Channel B	2025-02-03T12:00:00Z	3	8	500,000	25,000	2,500	00:05:30

---

4. Data Loading

The final transformed dataset was saved in a **TSV (Tab-Separated Values) file** named `trending_videos.tsv`. This format was chosen because:

- It preserves data integrity without issues caused by commas in titles.
- It is easily readable by Pandas, Excel, and other analytical tools.

File Saving Command:

```
df_trending.to_csv("trending_videos.tsv", sep="\t", index=False)
```

---

5. Key Findings and Insights

- The **most trending videos** tend to have high engagement (views, likes, and comments).
- **Time Since Published**: Many trending videos are **less than a week old**, indicating fresh content is favored.
- **Duration Trends**: Shorter videos (under 15 minutes) tend to dominate the trending list.
- **Channel Popularity**: Some channels consistently appear in the trending list, highlighting their strong audience engagement.

---

6. Limitations and Future Improvements

6.1 Limitations

- **API Quota Limitations**: Only a limited number of requests can be made daily.
- **Incomplete Data**: Some videos do not provide public statistics (e.g., hidden likes or comments).
- **Regional Trends**: Results are limited to the specified `regionCode` (e.g., India □□), which may not reflect global trends.

6.2 Future Improvements

- **Automate Daily Data Collection**: Schedule a script to collect trending videos **daily** for time-series analysis.
- **Expand to Multiple Regions**: Fetch and compare trending videos across different countries.
- **Sentiment Analysis**: Analyze video titles and comments to understand audience reactions.
- **Machine Learning Recommendations**: Build a model to predict potential trending videos based on historical data.

---

7. Conclusion

This project successfully implemented an **ETL pipeline** to extract, transform, and store trending YouTube videos. The structured dataset can now be used for further analysis, trend predictions, and insights into content engagement patterns.

Future improvements will focus on automating data collection and applying analytical techniques to derive deeper insights.

For any queries or improvements, please reach out!

End of Report