YouTube Data Collection Using Parallel Processing

Joseph Kready, Shishila Awung Shimray, Muhammad Nihal Hussain,

Nitin Agarwal
Department of Information Science
University of Arkansas at Little Rock (UALR)
Little Rock, USA
{jkready, sxawungshim, mnhussain, nxagarwal}@ualr.edu

Agenda

- Motivation
- Data Collection Methodology
- Function Overview
- Performance
- Conclusion
- Future Work

Introduction/Motivation/Goal

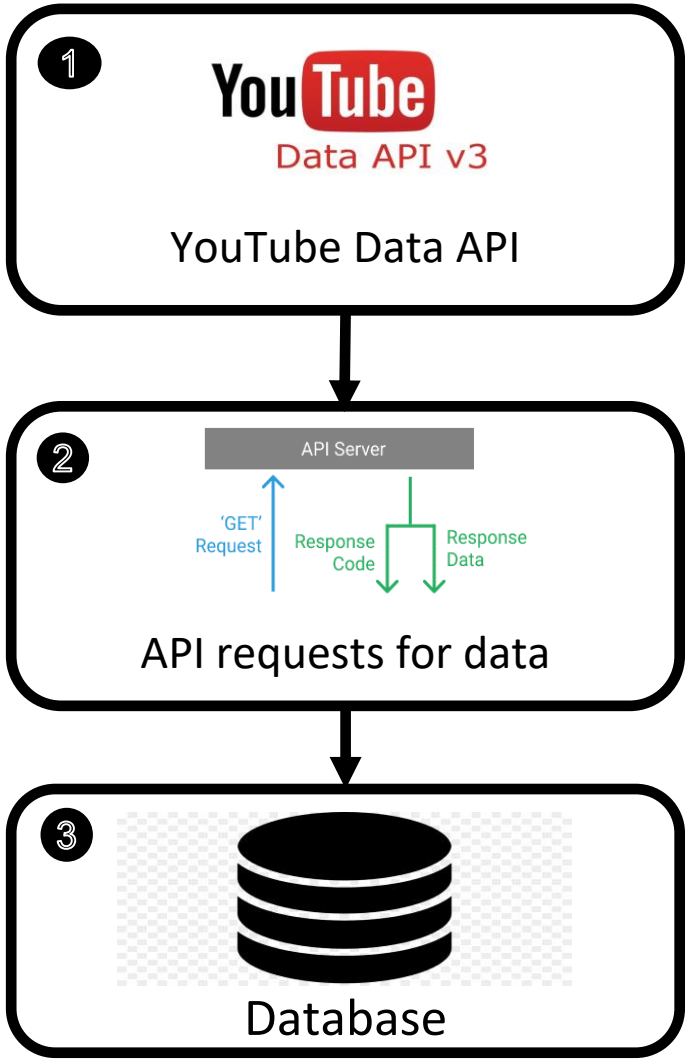
- YouTube is 2nd largest social media platform
- 10 Exabytes of data has been generated by YouTube
- Challenges of YouTube Analysis–
  - Slow sequential processing of API requests
  - API key daily usage limits

Data Collection Methodology

1. Obtain a YouTube Data API key

2. Develop a function to submit & process YouTube Data API requests

3. Store data for analysis



Function Overview

Looping through content IDs sequentially, making API requests one at a time

```
## - Single Process - ##
def single_process_video(video_ids):
    for video_id in video_ids:
        process_video(video_id)
```

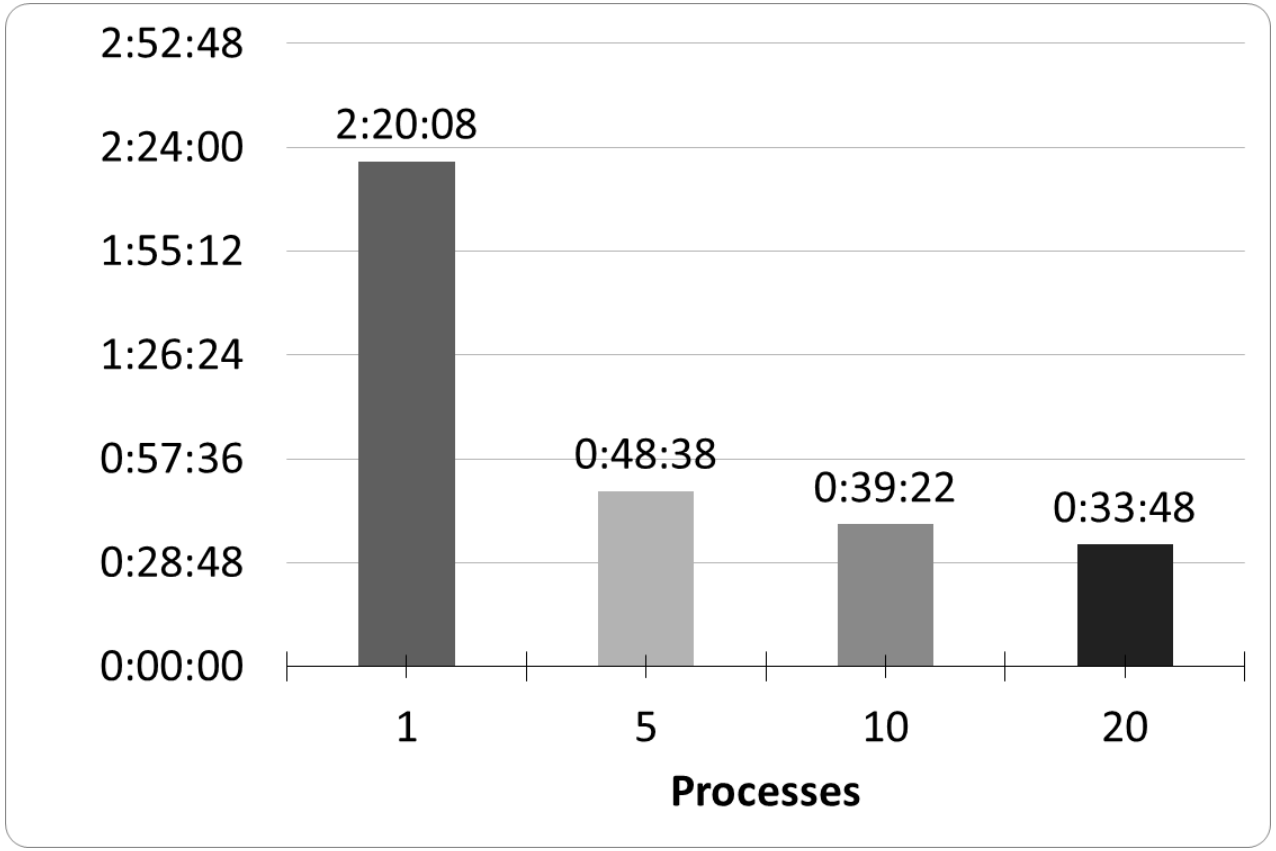
Function Overview

Splitting the Content IDs between 5 Nodes, making API requests in parallel

```
## - Parallel Process - ##
from pathos.multiprocessing import ProcessPool as Pool

def parallel_process_video(video_ids):
    #Creating a processing pool of 5 processes
    process_pool = Pool(nodes=5)

    #Mapping each video_id onto the process_video function
    process_pool.uimap(process_video, video_ids)

    process_pool.join()
    process_pool.close()
```

Performance

- Based on data processing times for FPSRussia channel
- A 400% decrease in processing time
- Biggest improvements from 1-5 processes



Conclusion/Future Work

- Parallelization of YouTube data collection dramatically decreases processing time
  - I/O bottlenecks are distributed across multiple processes
  - CPU can switch between processes while awaiting an API response

- Parallelized API requests can be used on other social media sites
  - Twitter
  - Reddit

Acknowledgments

This research is funded in part by the

- U.S. National Science Foundation,
- U.S. Office of Naval Research,
- U.S. Air Force Research Lab,
- U.S. Army Research Office,
- U.S. Defense Advanced Research Projects Agency, and
- Jerry L. Maulden/Entergy Fund at the UA-Little Rock.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations. The researchers gratefully acknowledge the support.

Thank You

For question/comments please email:

jkready@ualr.edu

nxagarwal@ualr.edu