Distributed Media Processing ft. Hadoop

Group 6

Intro

- Media transcoding is 'easy'
- Media transcoding is necessary
- Media transcoding can be horizontally scaled

Applications: Netflix, YouTube, Zoom, etc.

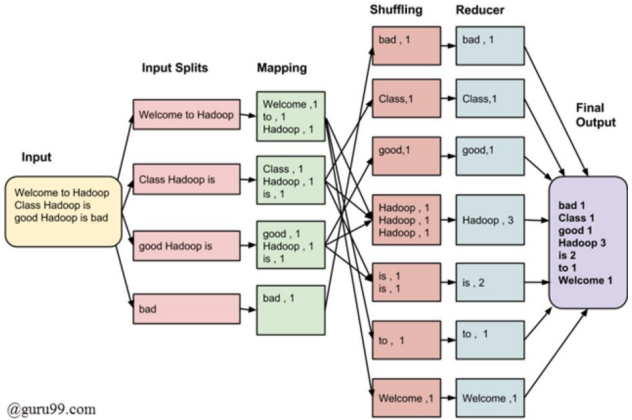
Architecture

The media downscaling operation consists of 3 steps:

1. Splitting up source video into chunks of roughly equal size (IO bound).
2. Applying the transformation to each chunk independently (CPU bound).
3. Merging the result chunks back into one file in the correct order (IO bound).

This plus Hadoop:

EMR is used (*)



Implementation

Python wrapper over Hadoop Streaming API

Uses HDFS CLI to interact with HDFS

Offloads actual video processing to FFMPEG.

Results

Curious results at first glance…

Then again, maybe not.

Media type	Duration for standalone mode	Duration for distributed mode
4K clip (30s)	15s	24s
1080p movie (2 hours)	56min	24min

Future scope

- GPU acceleration
- File streaming
- Use native Java