Research & Discovery at Scale: AI / ML & Cloud Services

Jim Duran, Director

Vanderbilt Television News Archive

VANDERBILT UNIVERSITY
Jean *and* Alexander Heard Libraries
Vanderbilt Television News Archive

James Pilkington, Administrator 1970 - 1986

About the Archive

- Founded in 1968
- A division of the Vanderbilt University Libraries
- A collection of TV news recordings
- AND metadata catalog
- Primary User Groups:
  - Academic Scholarship
  - Film and Media Industry
  - Open to the public

Demo Video:
Advanced Computing & Society

Challenges:

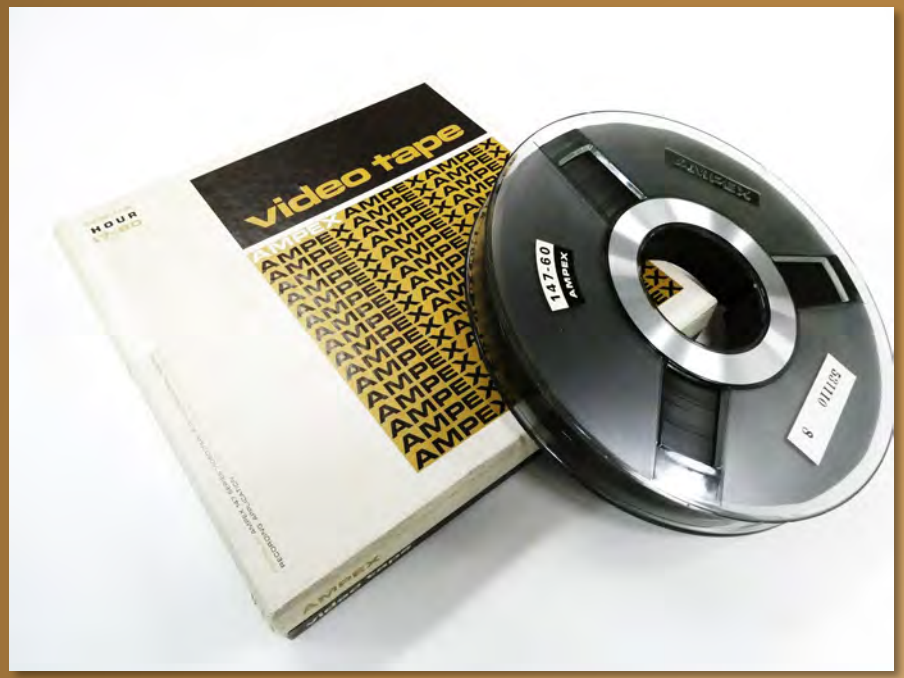
- Storage
- Video Flow
- Transcripts
- Description

Challenge: Storage

In 2018, The TV News Archive included 175 TB of archival video storage and added 7 TB every year. The department needed sustainable and cost-efficient storage.

1st transfer project: 1-inch open reel to 3/4

- AMPEX 1-inch Open Reel

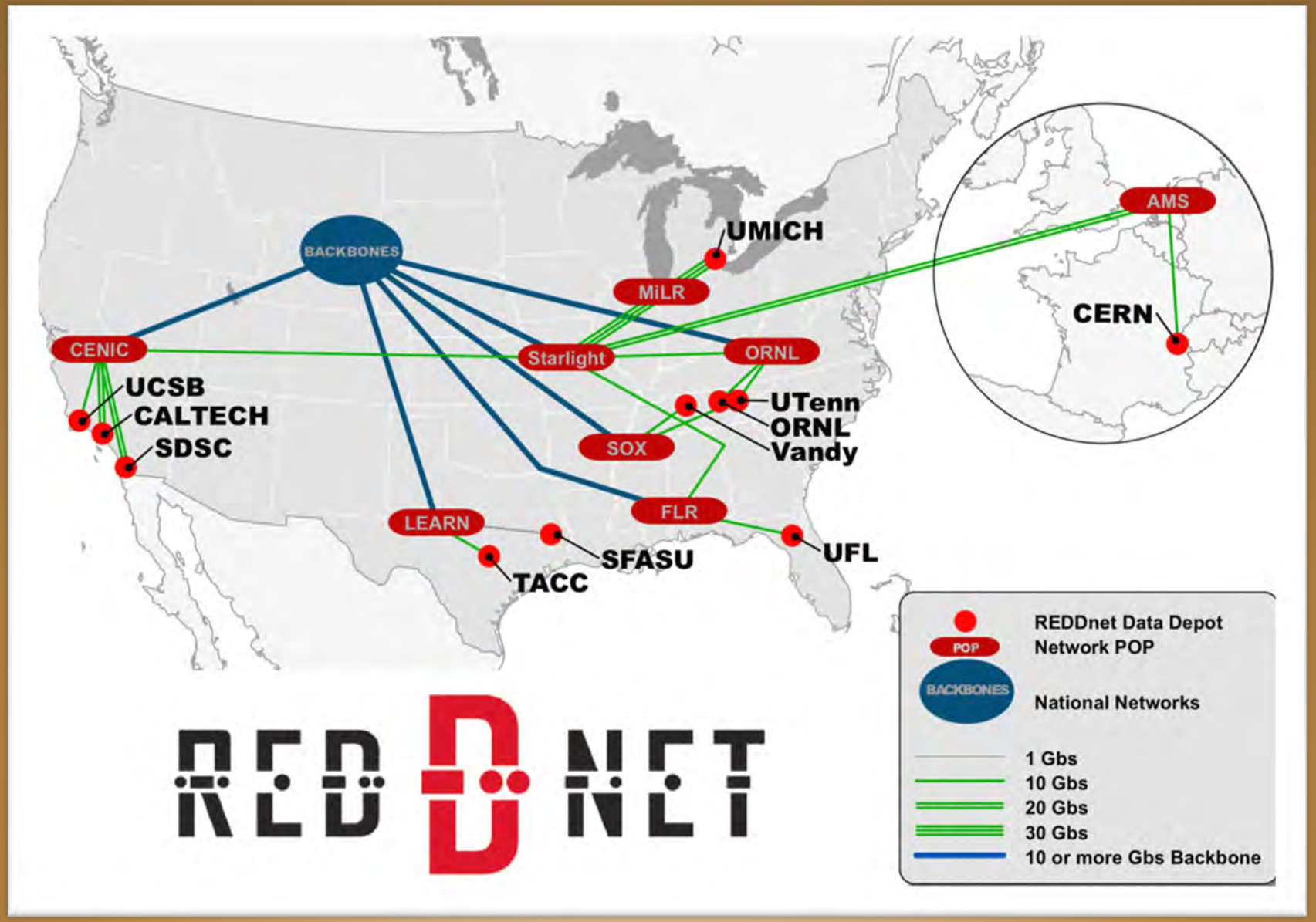


- U-MATIC ¾ Video Cassette

Second Transfer Project: *Digitization*

- Over 45,000 tapes
- Six years to finish
- Up to 10 stations running simultaneously
- Two shifts
- $1 Million Grant

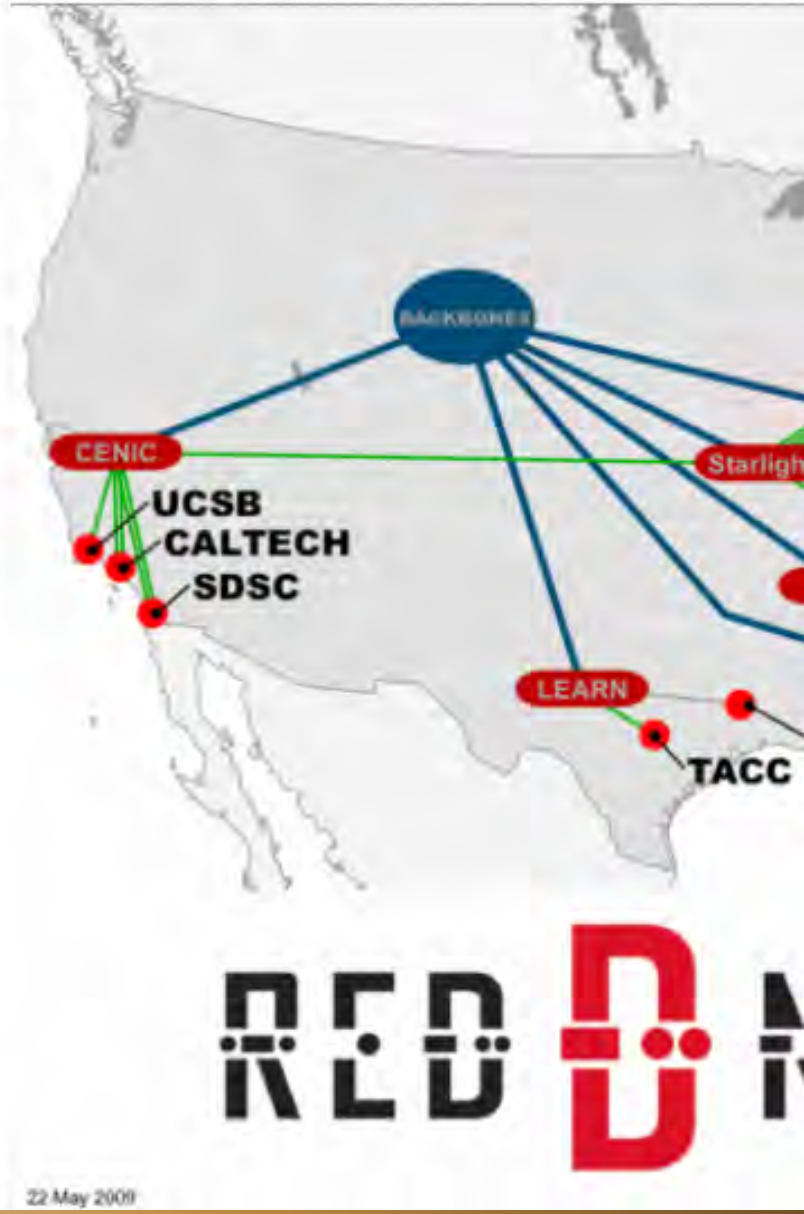
Storage Solutions and Requirements



REDDnet: Enabling Data Intensive Science in the Wide Area

REDDnet (Research and Education Data Depot network) is an NSF-funded infrastructure project designed to provide a large distributed storage facility for data intensive collaboration among the nation's researchers and educators in a wide variety of application areas. Its mission is to provide "working storage" to help manage the logistics of moving and staging large amounts of data in the wide area network, e.g. among collaborating researchers who are either trying to move data from one collaborator (person or institution) to another or who want share large data sets for limited periods of time (ranging from a few hours to a few months) while they work on it. REDDnet is not designed or intended to be a replacement for reliable archival or long term personal storage and users must make separate arrangements to insure that the data they are sharing via REDDnet's "best effort" storage is also preserved independently with stronger guarantees.

One example comes from the CMS collaboration, a high energy physics experiment that will be taking data soon at the Large Hadron Collider (LHC) at



AWS S3
Glacier Deep Archive

Started with new files written directly to AWS

Uploaded the 175 TB archive

- Could not use Snowball
- Set to Glacier Deep Archive

Proxy File

- More compression
- ¼ file size
- Intelligent Tiering for fast access

Challenge: Video Flow

The TV News Archive requires a highly reliable recording system that records the nightly news on five channels and maintains a 40-day buffer of television to monitor breaking news events.

Automated Recording

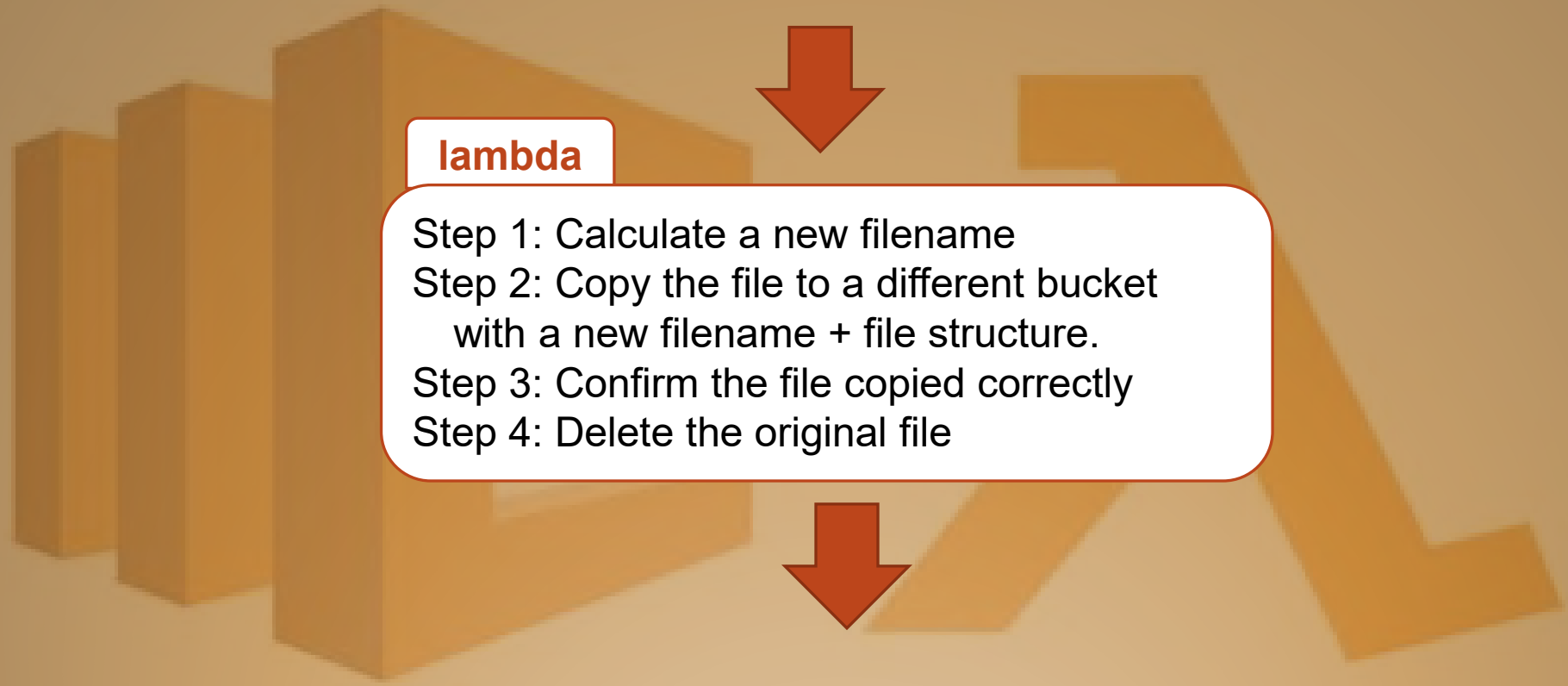
Lambda

Runs code without servers or clusters

1 million requests per month for free

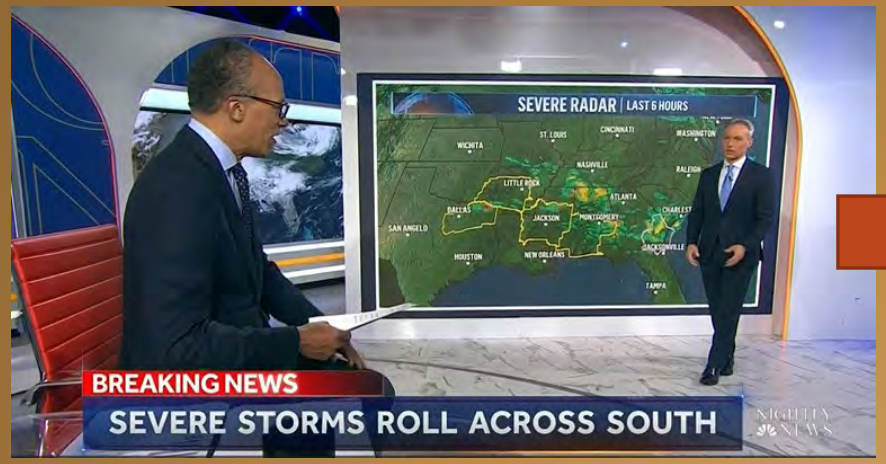
Executes a snippet of code

```
New File = "VID_20221207T082900.mp4"
```




```
Different Name & Location = "2022/12/VID20221207.mp4"
```

MediaConvert



Type: Primary Archive Version
Bitrate: 6 Mbps

Type: Proxy Access Version
Bitrate: 3 Mbps
Network, Date, Timestamp

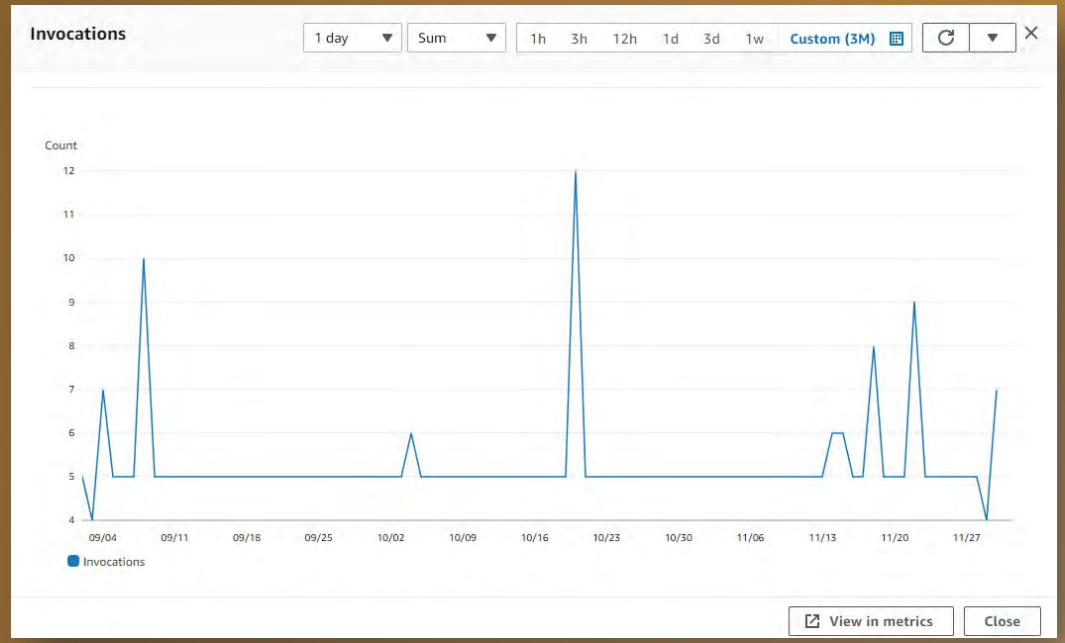
Lambda + MediaConvert

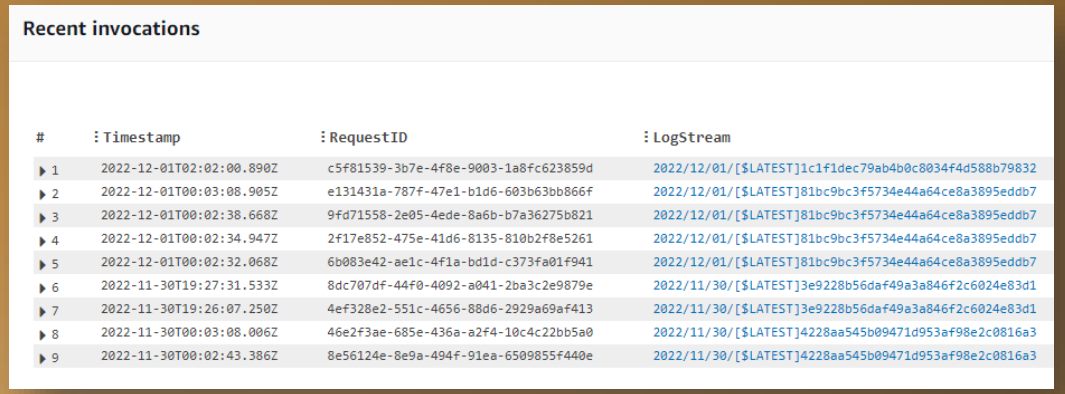
```
New File = "ABC_20190807T222901.mp4"
```

lambda

Step 1: Calculate Network, Date
Step 2: Convert Start time to Central Time
Step 3: Prepare settings for transcode job
Step 4: Start transcode job with watermark

Dashboards & Logs





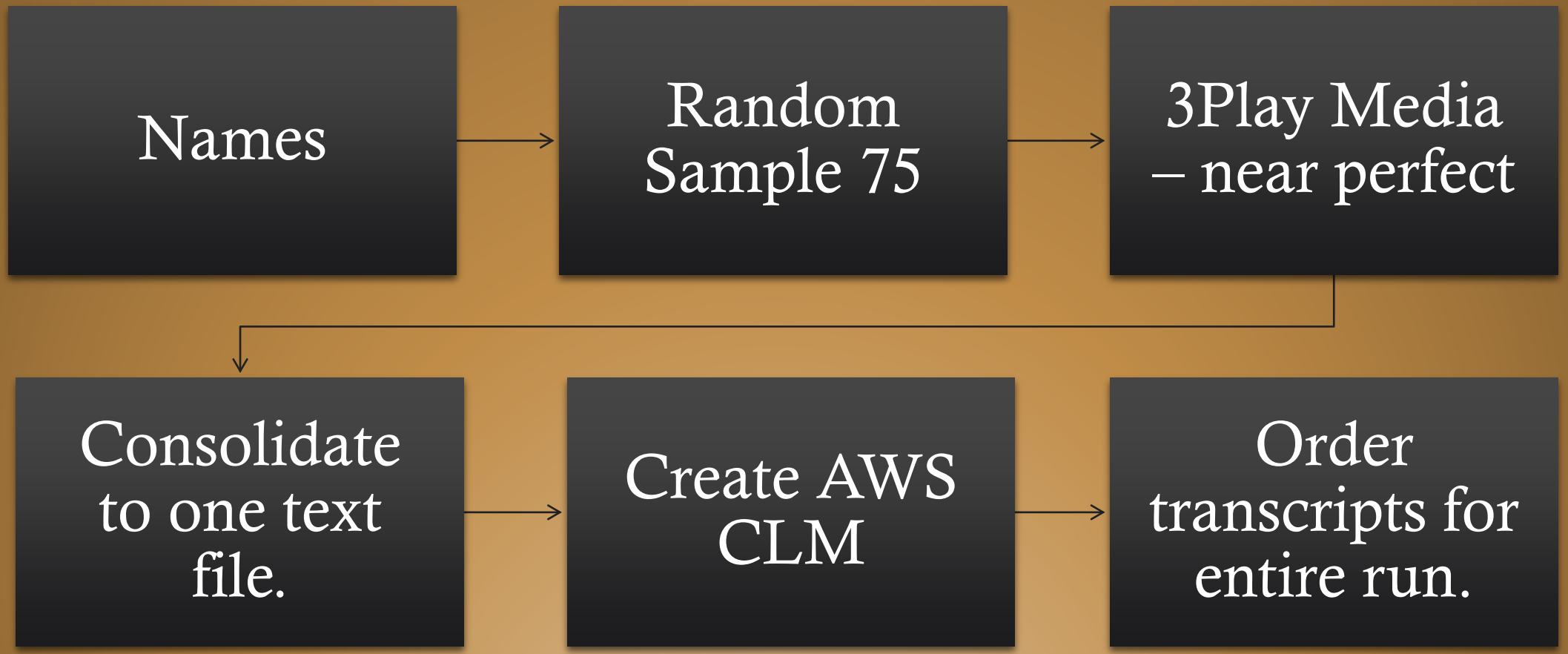
#	Timestamp	RequestID	LogStream
1	2022-12-01T02:02:00.890Z	c5f81539-3b7e-4f8e-9003-1a8fc623859d	2022/12/01/[$LATEST]1c1f1dec79ab4b0c8034f4d588b79832
2	2022-12-01T00:03:08.905Z	e131431a-787f-47e1-b1d6-603b63bb866f	2022/12/01/[$LATEST]81bc9bc3f5734e44a64ce8a3895eddb7
3	2022-12-01T00:02:38.668Z	9fd71558-2e05-4ede-8a6b-b7a36275b821	2022/12/01/[$LATEST]81bc9bc3f5734e44a64ce8a3895eddb7
4	2022-12-01T00:02:34.947Z	2f17e852-475e-41d6-8135-810b2f8e5261	2022/12/01/[$LATEST]81bc9bc3f5734e44a64ce8a3895eddb7
5	2022-12-01T00:02:32.068Z	6b083e42-ae1c-4f1a-bd1d-c373fa01f941	2022/12/01/[$LATEST]81bc9bc3f5734e44a64ce8a3895eddb7
6	2022-11-30T19:27:31.533Z	8dc707df-44f0-4092-a041-2ba3c2e9879e	2022/11/30/[$LATEST]3e9228b56daf49a3a846f2c6024e83d1
7	2022-11-30T19:26:07.250Z	4ef328e2-551c-4656-88d6-2929a69af413	2022/11/30/[$LATEST]3e9228b56daf49a3a846f2c6024e83d1
8	2022-11-30T00:03:08.006Z	46e2f3ae-685e-436a-a2f4-10c4c22bb5a0	2022/11/30/[$LATEST]4228aa545b09471d953af98e2c0816a3
9	2022-11-30T00:02:43.386Z	8e56124e-8e9a-494f-91ea-6509855f440e	2022/11/30/[$LATEST]4228aa545b09471d953af98e2c0816a3

Challenge: Transcripts & Captions

While summaries are helpful, many forms of research require a full account of what was said on TV. Closed captions are also essential for universal access.

62,000 hours transcribed in 2 months!

SPLIT THE COLLECTION BY PRESIDENTIAL ADMINISTRATION, THEN FOR EACH GROUP:



New Challenge: Descriptive Metadata

Describing archival material is currently the limiting factor in our capacity. Can we use large language models and machine learning to assist with descriptive metadata about news content?

Five Elements

Descriptive Metadata Field	Challenge	Possible Tools
Segment Start Time	Split a 30-minute news program into its segments: news story, commercials, other	Large Language Models (LLM) Computer Vision Machine Learning
Title	Summarize the news story into a brief title.	LLM
Summary	Summarize the news story in one paragraph.	LLM
People	Identify the people speaking in the news story.	LLM; NER; OCR;
Commercials	Identify the brands and products being advertised during the commercial breaks.	LLM; NER; OCR;

Promising Results: Title

Input:

"Turning now to the weather, with California already dealing with flooded roads and standing water, a second, more dangerous atmospheric river is bearing down on the state. The forecast. Let's bring in meteorologist Chris Warren from our partners at the Weather Channel. Good evening Chris. [Chris Warren, Meteorologist, The Weather Channel] Good evening Norah. Another strong storm moving into the West Coast. Going to bring a lot of rain. If Los Angeles ends up getting six inches of rain…"

Output:

California Braces for Dangerous Flooding

Promising Results: Commercial Products

CBS Evening News for November 22, 2023

First Commercial Block

10. Commercial
11. Commercial: Prevagen
12. Commercial: Swiffer power mop
13. Commercial
14. Commercial: Etsy's cyber sales event
15. Commercial
16. Commercial: Liberty
17. Commercial: Ashley

Second Commercial Block

20. Commercial: Living Wealth series
21. Commercial: Allstate
22. Commercial: Entyvio
23. Commercial: Downy Unstoppable
24. Commercial: Neutrogena Hydro Boost
25. Commercial: Weathertech
26. Commercial
27. Commercial
28. Commercial

Third Commercial Block

30. Commercial
31. Commercial: Listerine
32. Commercial: Consumer Cellular
33. Commercial: Purina One
34. Commercial
35. Commercial: Bounty
36. Commercial: Alka-Seltzer Plus
37. Commercial: Ashley
38. Commercial: Prudential
39. Commercial
40. Commercial: CBS Evening News

https://tvnews.vanderbilt.edu/siteindex/2023-11

Challenges Addressed

- Storage
- Video Flow
- Transcripts and Captions
- Generative Metadata

Thank you!

Jim Duran

Director
Television News Archive
Vanderbilt University Libraries

jim.duran@vanderbilt.edu

tvnews.vanderbilt.edu