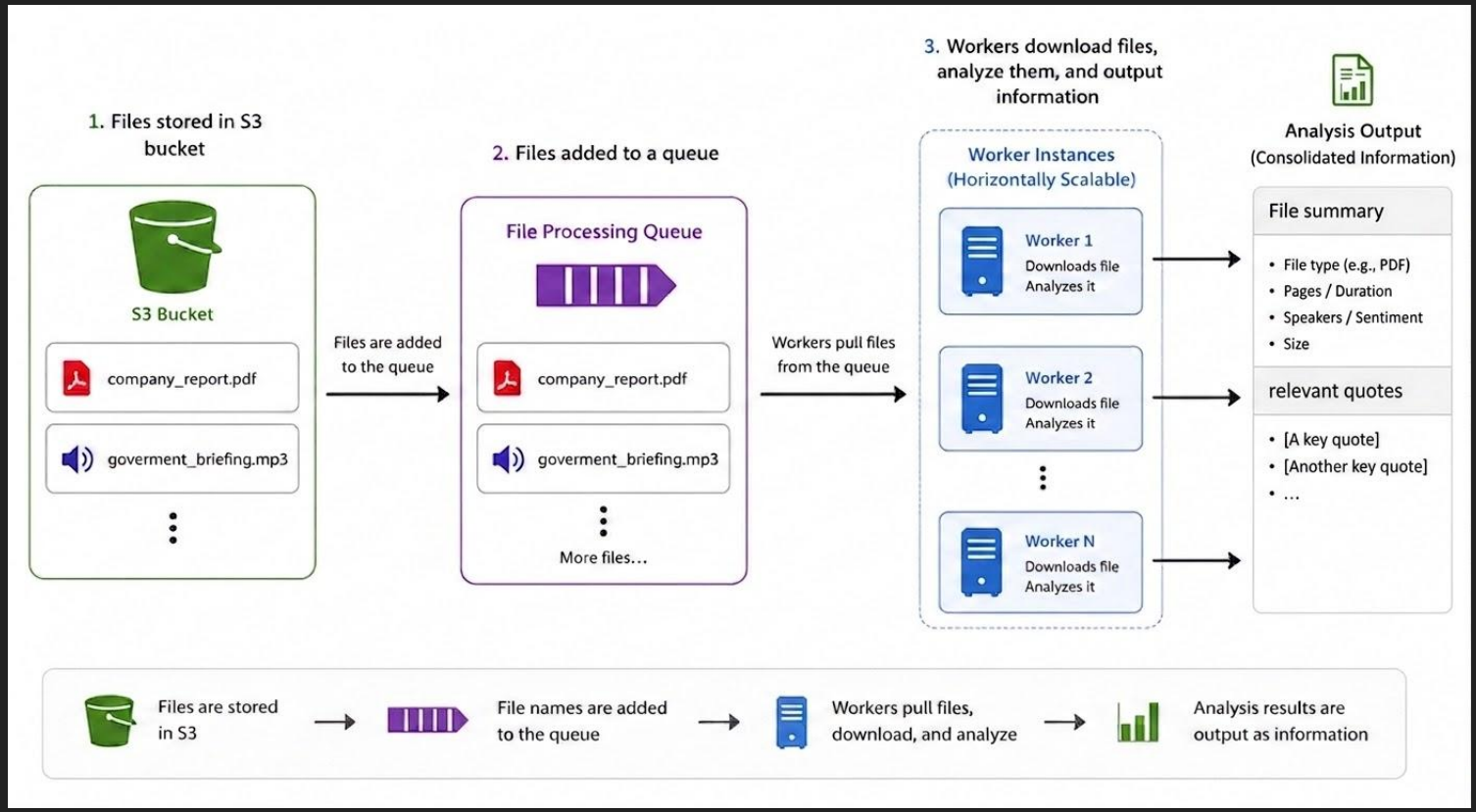
Using the cloud to rapidly analyse thousands of text and audio documents

Philip McMahon, The Guardian

https://github.com/philmcmahon/data-pipeline

philip.mcmahon@theguardian.com

What we're trying to do



This article is more than **2 months old**

Jeffrey Epstein's elite relationships visualised: the prince, the billionaire and the politicians

Guardian analysis of more than a million emails reveals financier's
powerful

3 seconds -17 secon
Postpone 5 Minutes Postpone 10 Minutes Skip Break

Day 2: th

Carmen Aguilar García, Michael Goodier, Philip McMahon, Ana Lucía González Paz, Paul Scruton *and* Paul Owen

Wed 18 Mar 2026 11.00 GMT

Share

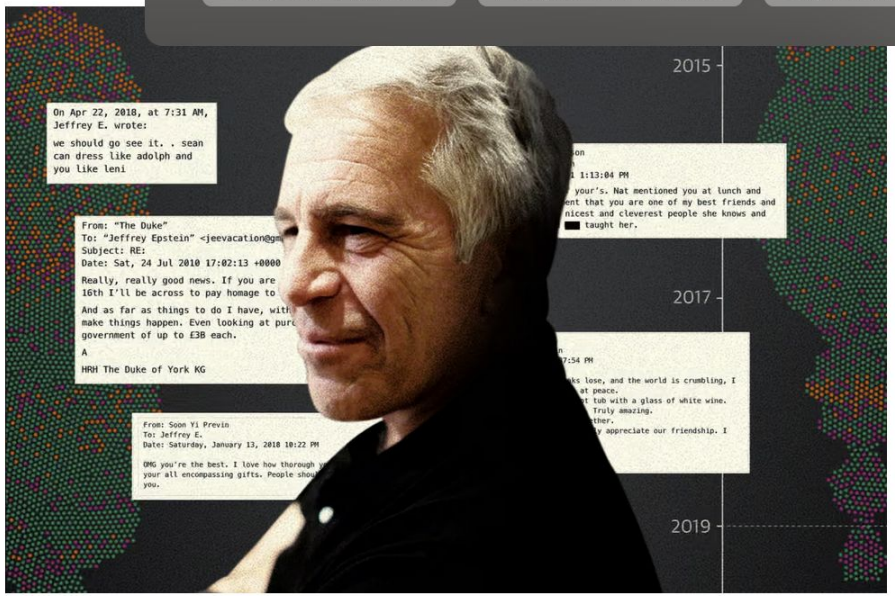


Illustration: Guardian Design/Getty

This article is more than **6 months old**

Influencers made millions pushing 'wild' births – now the Free Birth Society is linked to baby deaths around the world

Investigation by **Sirin Kale** *and* **Lucy Osborne**; *illustration by Laurie Avon*

Sat 22 Nov 2025 07.00 GMT

Share

A year-long investigation reveals how mothers lost children after being radicalised by uplifting podcast tales of births without midwives or doctors

Why bother?

- Look for existing tools first
- This technique can help save money/time
- Can have control over compromises made of cost/output quality/time

Cost

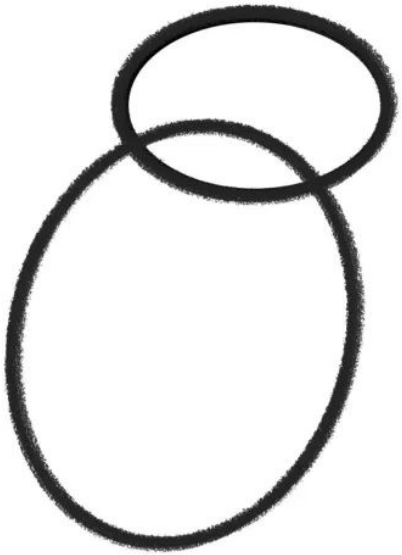
- Around $200 to extract text from 2 million epstein files
- Important to monitor
- Instances capable of doing transcription/LLM work cost $0.50/hour = $12/day

Cloud platforms

tinyurl.com/dh-data-pipeline

(redirects to github.com/philmcmahon/data-pipeline)

What we’re trying to learn/not learn



Step 1: Draw some circles

Step 2: Draw the rest of the owl!

Show passwords slide

Creating the infrastructure

- Source data bucket already exists
- Everyone needs to create their own:
  - Queue (SQS)
  - Worker pool (EC2 Autoscaling Group)
  - Output bucket

- Define infrastructure you need in a file that can be version controlled
- Works with any cloud provider
- `workers.tf` is the most of interest for this workshop

Transcription

- We are using https://github.com/m-bain/whisperX which needs a GPU to run efficiently.
  - In our experience 11x faster transcription time than the length of the file
- An alternative which doesn't require a GPU is https://github.com/ggml-org/whisper.cpp - particularly fast on MacOS
- Nvidia Parakeet is an even faster model - great for real time transcription (but a bit more difficult to set up)

OCR

- OcrMyPDF the guardian's favourite open source model
  - Used here alongside 'pdftotext' to get the raw text out at the end
  - Used for epstein files
  - Know your flags! https://ocrmypdf.readthedocs.io/en/latest/cookbook.html#redo-existing-ocr
- More recently some ones that use a GPU and 'fancy models'
  - https://docling-project.github.io/
  - https://github.com/PADDLEPADDLE/PADDLEOCR
  - Diagram support
  - Some code in the repo to try this out

Running prompts

- We used AWS Bedrock with an Anthropic model for Free Birth project
- Running it on your own instance is more effort but can be cheaper
- Allows you to run in a totally offline environment
- The server we use - vllm - is OpenAI API compatible, so it can easily be swapped out for an external API

Working with a GPU

- Loading models in/out of GPU is expensive
- So need to batch similar jobs (don’t interleave transcription with prompting)
- …or use a bigger instance

Thanks

philip.mcmahon@theguardian.com