M4 discussion

- **Look at your presentation time slots and resolve conflicts asap**
- Recap ETL techniques
- Common ETL techniques
    - FTP files
    - Web API
    - Distributed message queue - widely used today
    - Web scraping - not a resilient ETL

- Motivation for M4: You should think about which techniques make the most sense based on the given inputs and desired output of the ETL.
- Look at the Cinemalytics dataset - Collection of bollywood movies and music. Ingest this data into specific tables in the IMDB database.
- There's a version on S3 which doesn't have the headers
- Favour spark when copy commands choke. But here, the data size is small - You may go ahead with Postgres. You can also do a hybrid of ETL techniques.
- Organize your files based on the table you're populating.