# Personalizing Netflix with Streaming datasets

Shriya Arora

Senior Data Engineer
Personalization Analytics

@shriyarora

NETFLIX

## What is this talk about ?

- Helping you decide if a streaming pipeline fits your ETL problem
- If it does, how to make a decision on what streaming solution to pick

## What is this NOT talk about ?

- X streaming engine is the BEST, go use that one!
- Batch is dead, must stream everything!

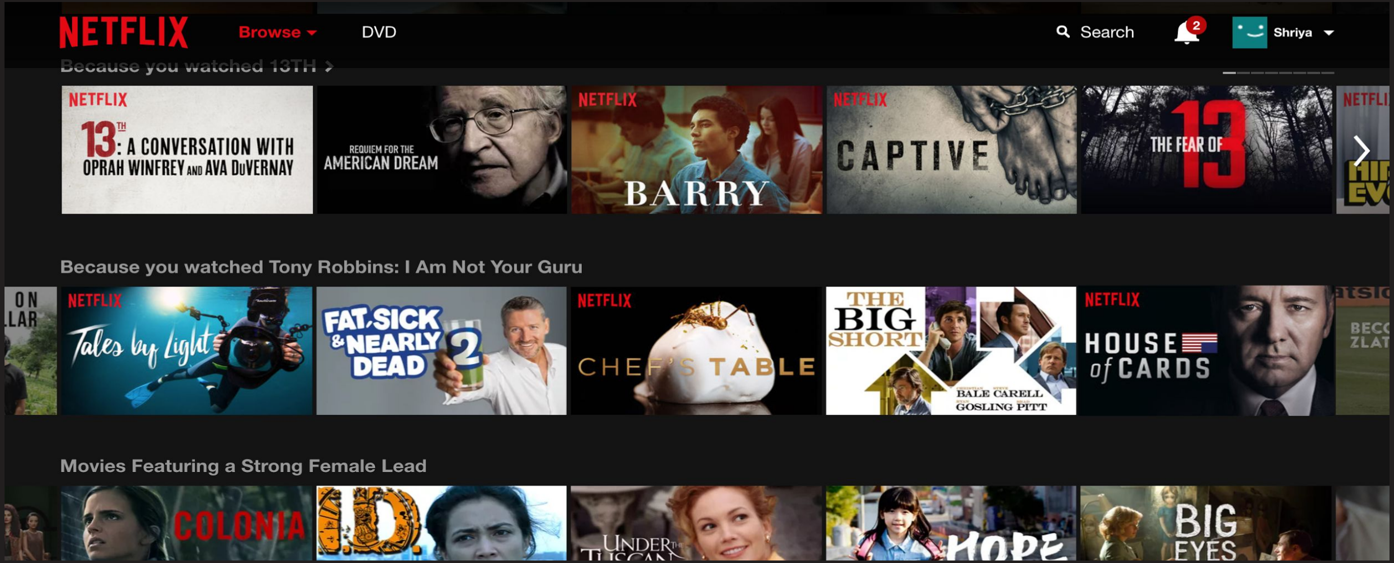
NETFLIX

# What is Netflix's Mission?

Entertaining you by allowing you to stream content anywhere, anytime

# What is Netflix's Mission?

## Entertaining you by allowing you to stream personalized content anywhere, anytime


NETFLIX
Browse
DVD
Search
2
Shriya
Because you watched 13TH >
13TH: A CONVERSATION WITH OPRAH WINFREY AND AVA DuVERNAY
REQUIEM FOR THE AMERICAN DREAM
BARRY
CAPTIVE
THE FEAR OF 13
Because you watched Tony Robbins: I Am Not Your Guru
Tales by Light
FAT, SICK & NEARLY DEAD 2
CHEF'S TABLE
THE BIG SHORT
BALE CARELL GOSLING PITT
HOUSE of CARDS
Movies Featuring a Strong Female Lead
COLONIA
I.D.
UNDER THE TUSCAN
HOPE
BIG EYES

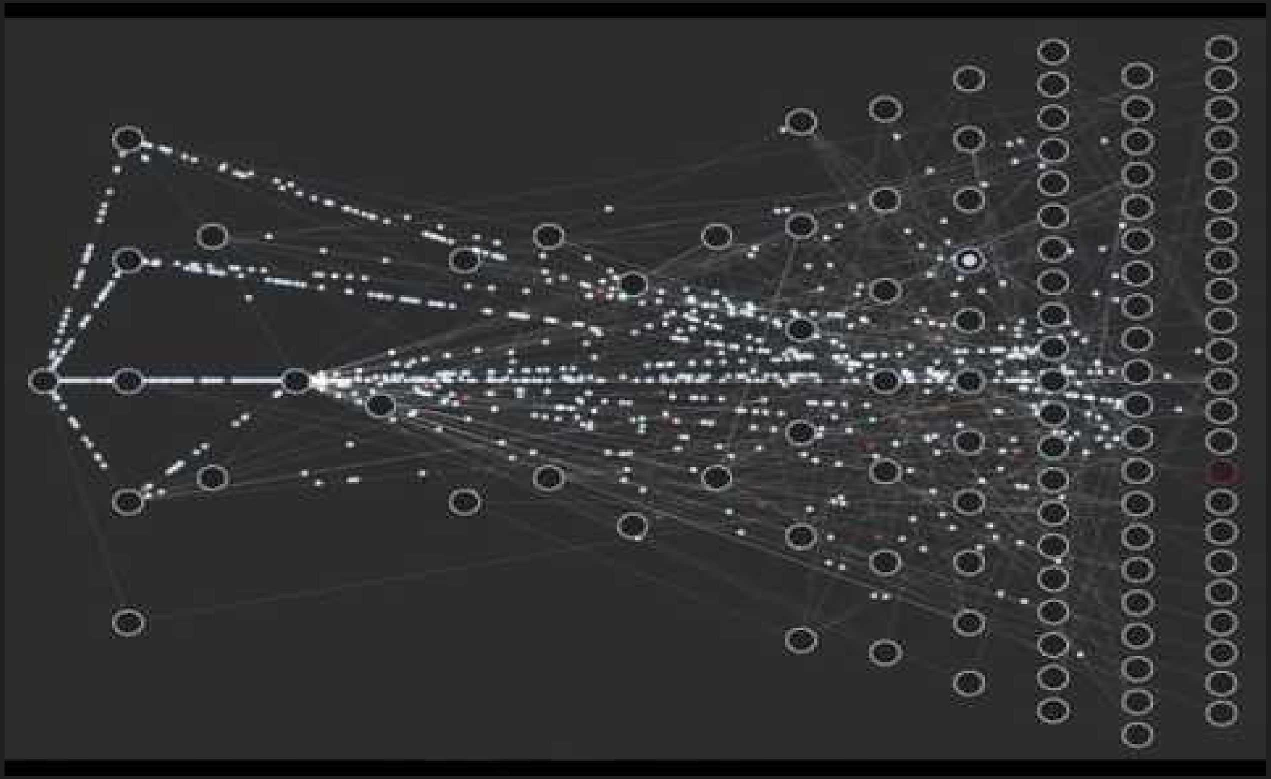
NETFLIX

# How much data do we process to have a personalized Netflix for everyone?

- 100M+ active members
- 125M hours/ day
- 190 countries with unique catalogs
- 450B unique events/day
- 700+ Kafka topics

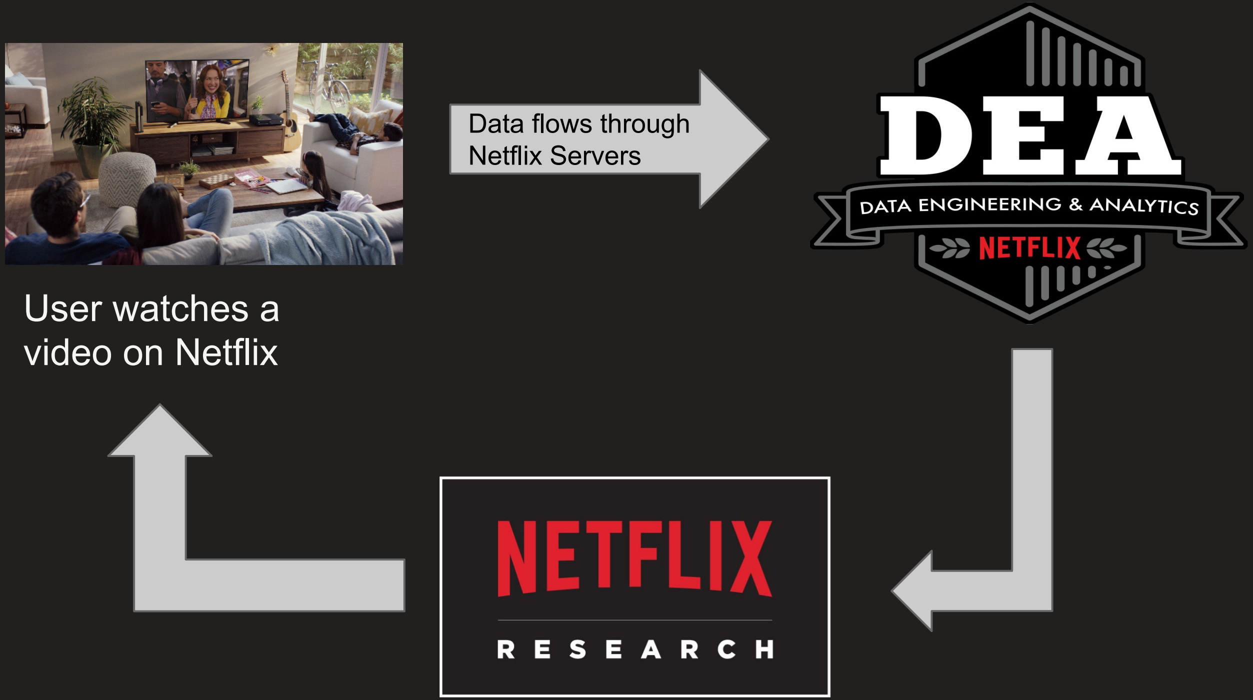
Image credit:http://www.bigwisdom.net/

# A series of Playback Events



NETFLIX

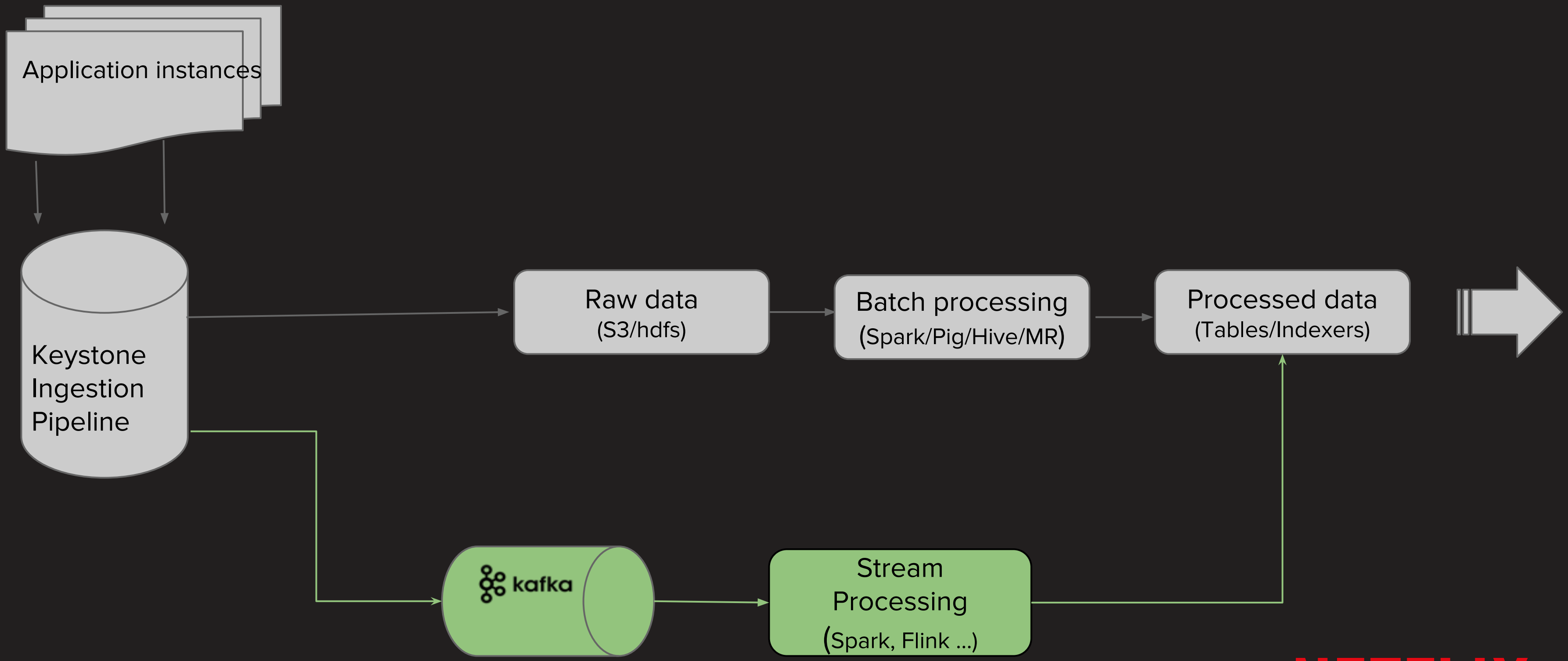
# DEA Personalization at a (very) high level



User watches a video on Netflix

Data flows through Netflix Servers

# Data Infrastructure

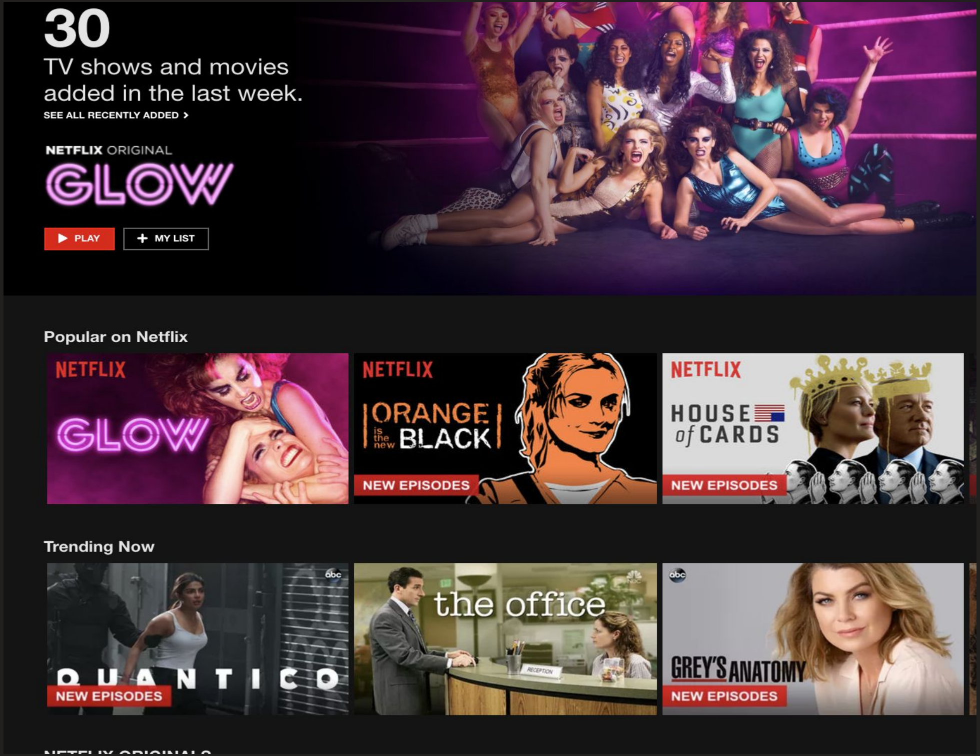

Application instances
Keystone Ingestion Pipeline
Raw data (S3/hdfs)
Batch processing (Spark/Pig/Hive/MR)
Processed data (Tables/Indexers)
kafka
Stream Processing (Spark, Flink ...)

NETFLIX

# Why have data later when you can have it now?

# Business wins

- Algorithms can be trained with the latest data



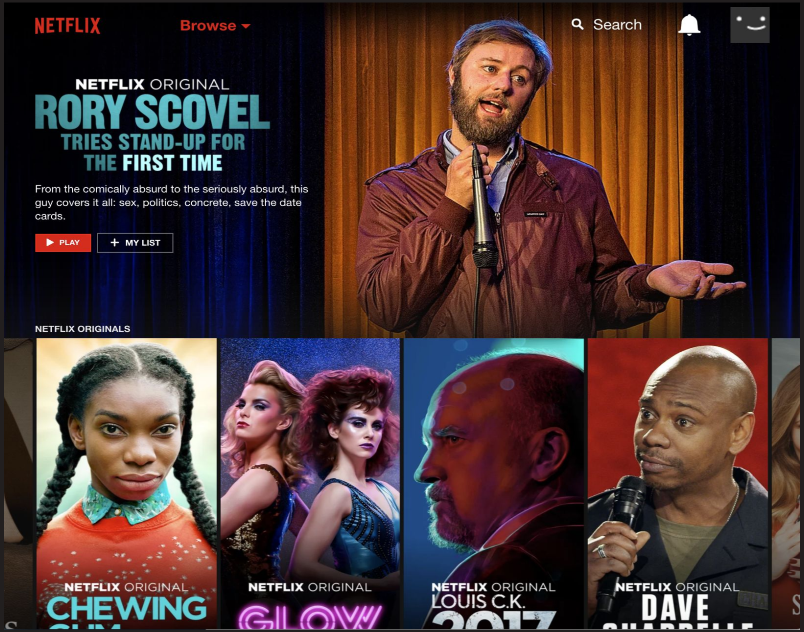
30

TV shows and movies added in the last week.

SEE ALL RECENTLY ADDED >

# Business wins

- Innovation in marketing of new launches



From the comically absurd to the seriously absurd, this guy covers it all: sex, politics, concrete, save the date cards.

- Creates opportunity for news kinds of algorithms

NETFLIX

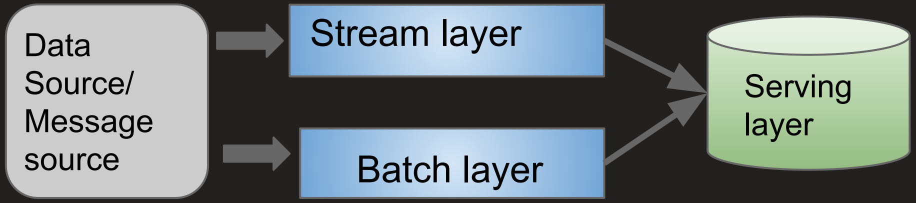
# Technical wins

- Save on storage costs
  - Raw data in its original form has to be persisted
- Faster turnaround time on error correction
  - Long-running batch jobs can incur significant delays when they fail
- Real-time auditing on key personalization metrics
- Integrate with other real-time systems
  - Additional infrastructure is required to make ‘online’ systems be available offline

# How to pick a Stream Processing Engine?

## Problem Scope/Requirements

- Event-based streaming or micro-batches?
- What features will be the most important for the problem?
- Do you want to implement Lambda?



# How to pick a Stream Processing Engine?

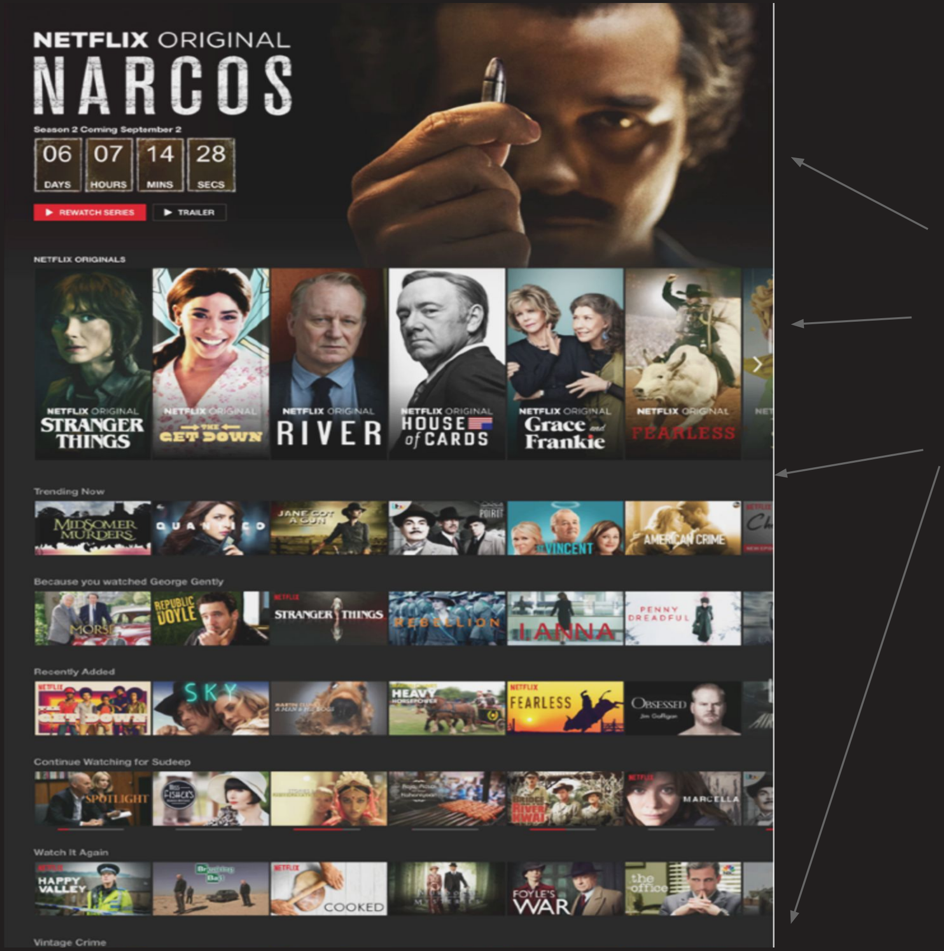
Existing Internal Technologies

- Infrastructure support: What are other teams using?
- ETL eco-system:  Will it fit in with the existing sources and sinks

What’s your team’s learning curve?

- What do you use for batch?
- What is the most fluent language of the team?

# Our problem: Source of Play / Source of Discovery



**Anatomy of a Netflix Homepage:**

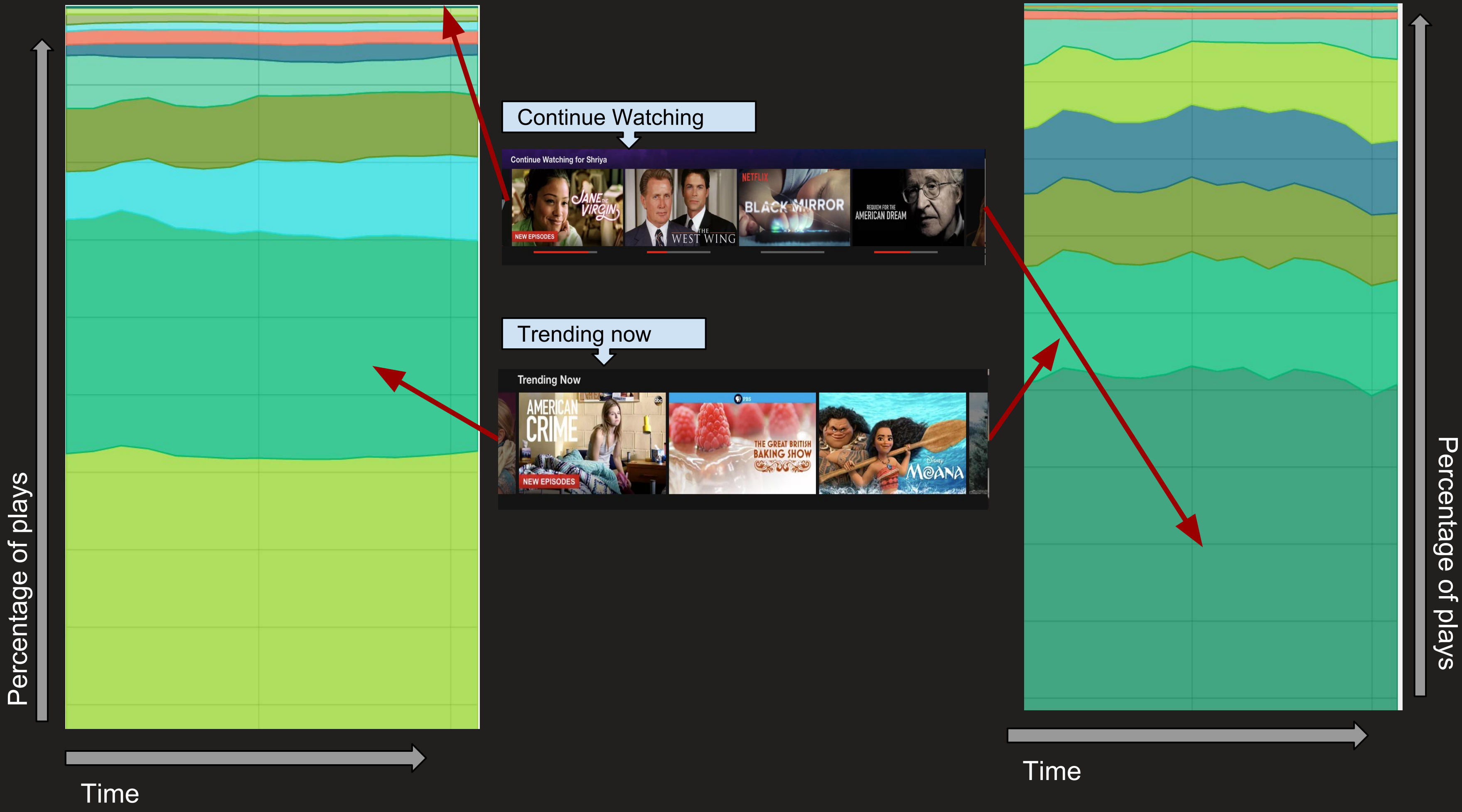
Billboard

Video Rankings (ordering of shows within a row)

Rows

## Source of Discovery

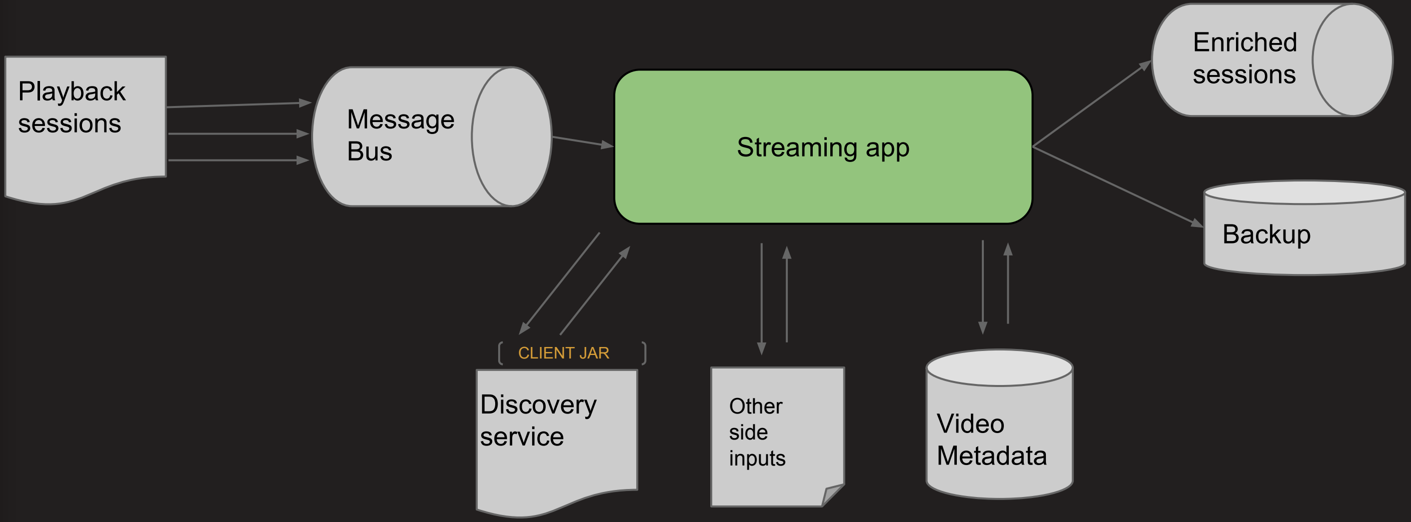
## Source of Play



## What we need to solve for Source of Discovery:

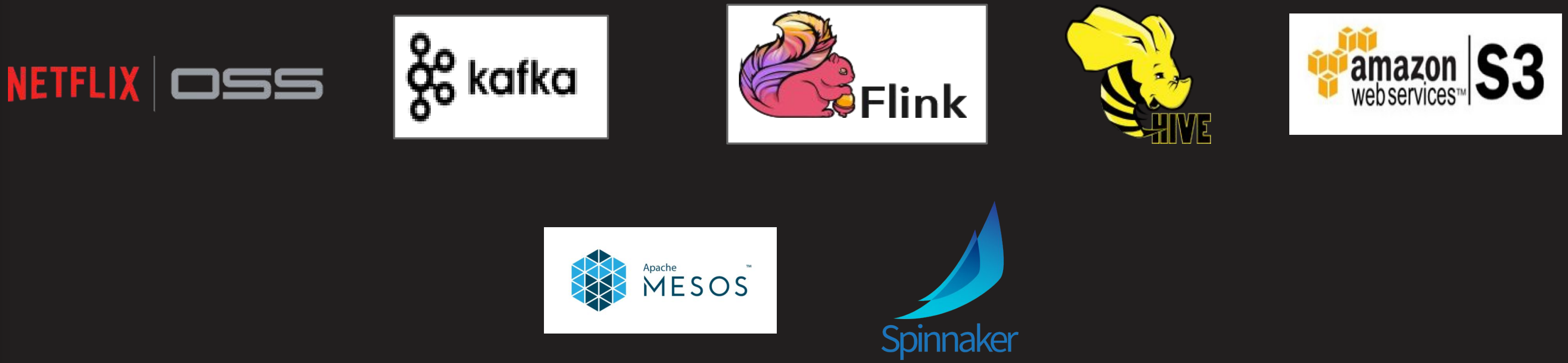
- High throughput
    - ~100M events/day
- Talk to live micro-services via thick clients
- Integrate with the Netflix platform eco-system
- Small State
- Allow for side inputs of slowly changing data

# Source-of-Discovery pipeline: Data Flow



NETFLIX

# Source-of-Discovery pipeline: Tech stack



By Source, Fair use, https://en.wikipedia.org/w/index.php?curid=47175041

## Getting streaming ETL to work

- Getting Data from Live sources
  - Every event (session) enriched with attributes from past history
  - Making a call to the micro-service via a thick client

- Side inputs
  - Get metadata about shows from the content service
  - Slowly changing data, optimize to call less frequently

- Dependency Isolation
  - Shading jars is fun *(said no one ever)*

## Getting streaming ETL to work cont..

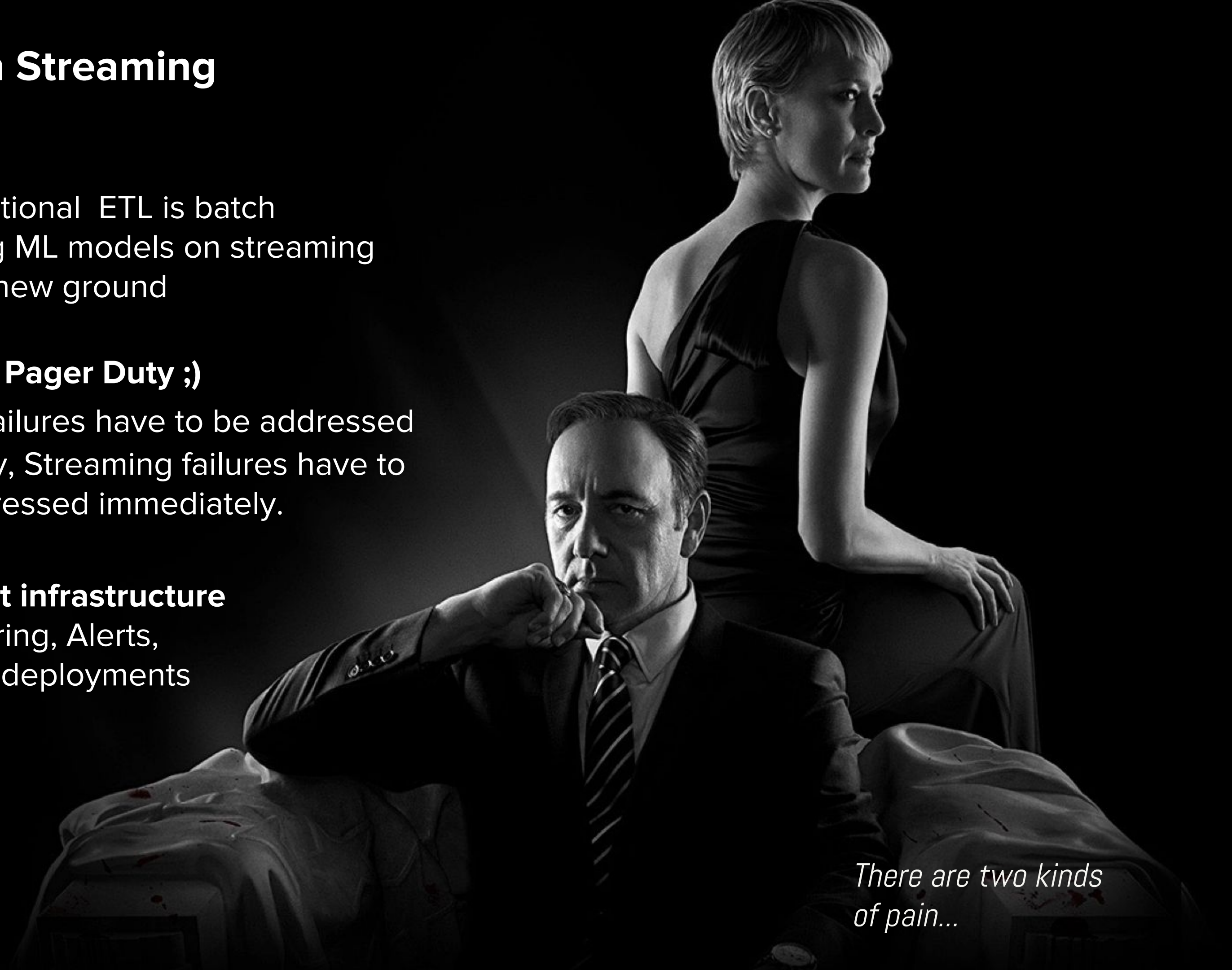
- Data Recovery
    - Kafka TTLs are aggressive
    - Raw data stored in HDFS for finite time for replay

- Out of order events
    - Late arriving data must be attributed correctly

- Increased Monitoring, Alerts
    - Because recovery is non-trivial, prevent data-loss

# Challenges with Streaming

- **Pioneer Tax**
  - Conventional  ETL is batch
  - Training ML models on streaming data is new ground
- **Outages and Pager Duty ;)**
  - Batch failures have to be addressed urgently, Streaming failures have to be addressed immediately.
- **Fault-tolerant infrastructure**
  - Monitoring, Alerts,
  - Rolling deployments



*There are two kinds of pain...*

# Questions?

Stay in touch!

@NetflixData