# Giới thiệu

(a) Câu chuyện netflix tầm (1 phút)

- Bối cảnh: Vào cuối những năm 1990, Netflix mới chỉ là một công ty khởi nghiệp chuyên cho thuê DVD qua đường bưu điện. Trong khi đó, Blockbuster là “ông lớn” thống trị thị trường cho thuê băng đĩa, với hàng nghìn cửa hàng trên khắp nước Mỹ.

- Netflix đề nghị bán mình:

+ Năm 2000, Netflix gặp nhiều khó khăn tài chính. Do mô hình kinh doanh chưa sinh lời, chi phí vận hành cao và sự cạnh tranh khốc liệt, hai nhà đồng sáng lập Reed Hastings và Marc Randolph đã đề nghị bán Netflix cho Blockbuster với giá khoảng 50 triệu USD.

+ Khi đó, CEO của Blockbuster là John Antioco đã từ chối thẳng thừng, cho rằng mô hình kinh doanh của Netflix không khả thi, và coi nó là một “ý tưởng nhỏ bé”.

- Hậu quả và sự đảo ngược ngoạn mục:

+ Sau khi bị từ chối, Netflix tiếp tục cải tiến mô hình kinh doanh, sau đó chuyển mình sang mô hình streaming (xem phim trực tuyến), trở thành một trong những công ty công nghệ và truyền thông lớn nhất thế giới.

+ Ngược lại, Blockbuster không thích nghi kịp với sự thay đổi trong thói quen người tiêu dùng và tuyên bố phá sản vào năm 2010.

(b) Điều gì làm một Netflix thành công

- Netflix chính thức "data-driven by design" từ năm 2006, và trở thành data-first trong mọi quyết định từ 2013 trở đi

- Netflix còn là data-driven company. Nhiều chiến dịch kinh doanh của doanh nghiệp đều dựa trên insights có được nhờ data analysis. Theo thống kê trên datap pipeline:

+ ~500 tỷ sự kiện và ~1,3 PB mỗi ngày

+ ~8 triệu sự kiện và ~24 GB mỗi giây trong giờ cao điểm

(nguồn: [Evolution of the Netflix Data Pipeline | by Netflix Technology Blog | Netflix TechBlog](https://netflixtechblog.com/evolution-of-the-netflix-data-pipeline-da246ca36905))

- Yếu tố công nghệ và kỹ thuật là không thể thiếu:

+ Có hệ thống đề xuất cá nhân hoá tốt: “75–80% nội dung người dùng xem trên Netflix đến từ các đề xuất được cá nhân hóa” Nguồn: https://www.businessinsider.com/netflixs-recommendation-engine-drives-75-of-viewership-2012-4 của báo business insider

+ Nắm bắt tốt nhu cầu thị trường để sản xuất phim: The Witcher (2007), Squid Game (2021), Wednesday (2022)

+ Tối ưu hoá trải nghiệm người dùng.

=> 3 yếu tố này đều có dự đóng góp quan trọng của dữ liệu.

# Vòng đời dữ liệu tại Netflix

A screenshot of a computer

AI-generated content may be incorrect.

# Nguồn dữ liệu

- Dữ liệu hành vi người dùng (<https://help.netflix.com/vi/node/100639>) :

+ Lịch sử xem: Các chương trình và phim bạn đã xem, thời lượng xem, và thời điểm xem trong ngày.

+ Tương tác: Đánh giá, lượt thích/không thích, và hành vi tìm kiếm, chia sẻ.

+ Hành vi người dùng tương tự: So sánh với người dùng có sở thích giống bạn.

+ Thiết bị và ngôn ngữ: Loại thiết bị sử dụng và ngôn ngữ ưa thích.

+ Thể loại yêu thích: Ai đang xem thể loại nào, mức độ thường xuyên.

+ Tình huống bỏ phim giữa chừng: Khi nào thì người xem ngừng xem và lý do có thể là gì.

- Phân tích xu hướng bên ngoài:

+ Mạng xã hội & diễn đàn: Theo dõi chủ đề đang "hot" trên Twitter, TikTok, YouTube, Reddit, Facebook Groups...

+ Google Trends: Xác định xu hướng tìm kiếm theo khu vực và thời gian.

+ Trang đánh giá: Thu thập phản hồi và xếp hạng từ IMDb, Rotten Tomatoes, Letterboxd.

+ Diễn đàn điện ảnh: Nắm bắt thảo luận, ý kiến từ cộng đồng yêu phim.

- Dữ liệu từ đối tác & thị trường:

+ Báo cáo nghiên cứu ngành: Từ Nielsen, Statista, McKinsey, PwC...

+ Thông tin từ chuỗi cung ứng nội dung: Nhà sản xuất, phân phối, rạp chiếu phim.

=> Dữ liệu đầu vào có đủ 3 loại dữ liệu

A screenshot of a white text box

AI-generated content may be incorrect.

## **Ingestion – Tiếp nhận và phân phối dữ liệu**

### **3.1. Luồng tiếp nhận dữ liệu**

Khi người dùng chọn một bộ phim, hệ thống frontend — được thiết kế để hiển thị nội dung trên web hoặc thiết bị di động — sẽ gửi một yêu cầu API đến backend. Backend xử lý các yêu cầu này và điều phối việc phân phối nội dung thông qua các dịch vụ streaming, đảm bảo video được phát mượt mà đến người dùng. Bên cạnh đó, dữ liệu được tạo ra từ các tương tác này sẽ được lưu trữ và xử lý nhằm cải thiện dịch vụ liên tục

Trong hệ thống data pipeline của Netflix, giai đoạn ingestion đóng vai trò quan trọng trong việc thu thập và điều phối dữ liệu từ các nguồn đầu vào đến các hệ thống phân tích. Dữ liệu sự kiện từ phía người dùng như hành vi xem (play, pause), đăng nhập hay tìm kiếm được gửi từ ứng dụng về backend thông qua API REST hoặc GraphQL.

Sau đó, các microservices chịu trách nhiệm phát sinh và đẩy các sự kiện này vào Kafka, thông qua các producer nội bộ hoặc Keystone Pipeline – một framework ingestion chuẩn hóa của riêng Netflix. Dữ liệu được gửi vào Fronting Kafka – cụm Kafka đầu tiên đóng vai trò làm bộ đệm trung gian. Từ đây, dữ liệu tiếp tục được định tuyến sang các Consumer Kafka clusters, phục vụ cho các hệ thống xử lý downstream như Apache Flink (streaming) hoặc Apache Spark (batch).

Cách tiếp cận này giúp Netflix đạt được khả năng ingestion hàng trăm tỷ sự kiện mỗi ngày, đồng thời tối ưu giữa tốc độ xử lý thời gian thực và khả năng phân tích chuyên sâu theo lô.

## **Data Transformation – Chuyển đổi dữ liệu tại Netflix**

### **4.1. Mục tiêu và vai trò**

Data Transformation là giai đoạn trung tâm trong chu trình xử lý dữ liệu tại Netflix, nhằm chuyển đổi dữ liệu thô — sinh ra từ các tương tác người dùng — thành dữ liệu sạch, có cấu trúc và giá trị sử dụng. Dữ liệu sau biến đổi sẽ phục vụ cho nhiều mục đích downstream như: dashboard giám sát, hệ thống đề xuất nội dung (recommendation), các thử nghiệm A/B, huấn luyện mô hình machine learning, và phân tích hiệu quả vận hành.

Quá trình này được triển khai dưới hai hình thức chính: **streaming (thời gian thực)** và **batch (theo lô định kỳ)**, tùy theo độ trễ chấp nhận được và mục tiêu phân tích cụ thể.

### **4.2. Streaming Processing – Xử lý dòng dữ liệu thời gian thực**

#### **a) Công cụ sử dụng**

Netflix triển khai xử lý streaming chủ yếu với hai nền tảng:

* **Apache Flink**: cho phép xử lý dòng dữ liệu phân tán với độ trễ thấp, khả năng mở rộng cao.
* **Mantis**: nền tảng nội bộ được thiết kế riêng cho quan sát (observability), cảnh báo và giám sát hệ thống.

#### **b) Các bước xử lý chính**

Dòng dữ liệu từ **Kafka** được xử lý theo chuỗi các bước sau:

* **Filter**: loại bỏ các sự kiện không cần thiết hoặc trùng lặp.
* **Enrichment**: gắn thêm thông tin phụ trợ (metadata) từ các nguồn như cơ sở dữ liệu nội dung.
* **Join**: kết hợp dữ liệu từ nhiều dòng (ví dụ: hành vi người dùng kết hợp với metadata phim).
* **Aggregation**: tổng hợp theo cửa sổ thời gian (ví dụ: số lượt xem mỗi 5 phút cho mỗi video).
* **Format**: chuẩn hóa và chuyển đổi định dạng dữ liệu đầu ra.

#### **c) Đầu ra và ứng dụng**

Sau xử lý, dữ liệu được ghi vào các hệ thống lưu trữ phục vụ cho các ứng dụng phản hồi tức thì:

* **Apache Druid**: hỗ trợ truy vấn phân tích thời gian thực.
* **Apache Iceberg**: lưu trữ dữ liệu có cấu trúc theo định dạng bảng.
* **Amazon S3**: lưu trữ dữ liệu gốc phục vụ các pipeline tiếp theo.

**Ví dụ thực tiễn**: Một pipeline Flink thu thập và phân tích số lần nhấn “pause” trong video để phát hiện các đoạn nội dung khiến người xem mất tập trung, từ đó hỗ trợ cải thiện chất lượng nội dung.

### **4.3. Batch Processing – Xử lý theo lô định kỳ (ETL)**

#### **a) Công cụ sử dụng**

Trong trường hợp không yêu cầu phản hồi tức thì, Netflix triển khai các pipeline xử lý theo lô (batch) với các công cụ chính:

* **Apache Spark**: xử lý dữ liệu lớn từ S3 hoặc Kafka theo batch.
* **Presto**: engine truy vấn SQL phân tán chạy trực tiếp trên data lake.
* **Hadoop**: lưu trữ và xử lý dữ liệu phân tán.
* **Meson**: hệ thống điều phối workflow dạng DAG (Directed Acyclic Graph).

#### **b) Quy trình xử lý**

Các bước phổ biến trong quy trình ETL bao gồm:

* **Cleaning**: loại bỏ dữ liệu lỗi, thiếu, trùng lặp.
* **Normalization**: chuẩn hóa schema, định dạng timestamp, định danh nội dung.
* **Join & Aggregate**: kết hợp dữ liệu từ logs người dùng, metadata phim, kết quả recommendation.
* **Advanced Transformation**: phân cụm người dùng, phân tích chuỗi hành vi, xây dựng đặc trưng đầu vào cho mô hình ML.

#### **c) Đầu ra và ứng dụng**

Sau khi xử lý, dữ liệu batch được ghi vào các hệ thống:

* **Apache Iceberg**: phục vụ truy vấn phân tích quy mô lớn.
* **Amazon Redshift**: hỗ trợ báo cáo BI và dashboard.
* **Các hệ thống BI** như **Tableau**: dùng cho phân tích trực quan và ra quyết định chiến lược.

### **4.4. Tổng kết**

Netflix kết hợp linh hoạt giữa **streaming processing** và **batch processing** để đạt được sự cân bằng giữa:

* **Phản ứng tức thì (real-time insight)** cho các nhu cầu vận hành và cá nhân hóa.
* **Phân tích sâu (deep analytics)** cho chiến lược dài hạn và ra quyết định dựa trên dữ liệu.

**Nguồn**:

* [Apache Flink at Netflix](https://flink.apache.org/2020/04/21/how-netflix-uses-apache-flink.html)
* Mantis: Netflix Observability Platform
* Spark ETL at Netflix
* Data Engineering at Netflix

**Nguồn**:

* Netflix’s EVCache
* [Netflix Search with Elastic](https://www.elastic.co/blog/netflix-uses-elasticsearch)

**VII. Serving**

🔹 **Hệ thống đề xuất cá nhân hoá tốt**

* Netflix sử dụng **machine learning models** huấn luyện trên data pipeline để đưa ra đề xuất phù hợp với từng người dùng.
* Hệ thống bao gồm nhiều tầng:
  + **Page Generation System** tạo homepage động cho từng user.
  + **Ranking Models** xếp hạng nội dung theo hành vi xem, thời gian, đánh giá, chủ đề quan tâm.
  + **Bandit Algorithms** để thử nghiệm nội dung mới dựa trên độ tương tác người dùng.

🔹 **Nắm bắt nhu cầu thị trường để sản xuất phim**

* Kết hợp dữ liệu nội bộ (hành vi người dùng) và dữ liệu bên ngoài (trends, diễn đàn, social media).
* 🧬 Data Science được dùng để:
  + Dự đoán xu hướng thể loại, chủ đề hấp dẫn.
  + Xác định thời điểm phát hành tối ưu.
  + Chọn diễn viên/đạo diễn phù hợp với phân khúc khán giả.

🔹 **Tối ưu hóa trải nghiệm người dùng (UX)**

* 📱 Giao diện được cá nhân hóa theo thiết bị, vị trí, và sở thích.
* 🧪 A/B Testing liên tục để cải tiến UI/UX:
  + Vị trí nút "Play", thumbnail động, preview ngắn.
* ⚙️ Hệ thống theo dõi lỗi và performance real-time giúp tối ưu độ trễ khi stream.
* 💡 Netflix có hệ thống observability tinh vi (Mantis) giúp phát hiện sớm các bất thường UX và cải thiện ngay lập tức.

**VIII. Vai trò của AWS trong hệ thống dữ liệu của Netflix**

1. **Lý do chọn AWS**

Netflix bắt đầu chuyển toàn bộ hệ thống sang AWS từ năm 2008 và hoàn tất vào năm 2016. Đây là một trong những case study thành công nhất về “cloud migration” trong ngành công nghệ.

Netflix chọn AWS vì:

* Khả năng mở rộng linh hoạt (elasticity): AWS cho phép Netflix tự động mở rộng tài nguyên theo nhu cầu từng thời điểm (đặc biệt trong giờ cao điểm hoặc phát hành phim hot).
* Tính sẵn sàng cao (high availability): AWS cung cấp kiến trúc đa vùng (multi-region), giúp Netflix duy trì uptime gần như tuyệt đối.
* Hệ sinh thái dịch vụ phong phú: từ lưu trữ, xử lý dữ liệu lớn, máy học, đến bảo mật đều có thể tích hợp ngay mà không cần tự xây.

1. **Chiến lược cloud của Netflix: “Everything on AWS” nhưng vẫn có tự chủ**

Dù sử dụng rất nhiều dịch vụ của AWS, Netflix không “phụ thuộc hoàn toàn”, mà:

* Phát triển hệ thống riêng: như *Mantis* (quan sát), *EVCache* (cache), *Atlas* (monitoring), *Zuul* (API gateway).
* Quản lý cấu hình và CI/CD riêng biệt: thông qua Spinnaker (do chính Netflix phát triển, nay là mã nguồn mở).

Điều này giúp Netflix:

Tận dụng hạ tầng mạnh mẽ của AWS,

Nhưng vẫn giữ quyền kiểm soát kiến trúc phần mềm và tránh bị “vendor lock-in” toàn phần.

***“Netflix là một ví dụ điển hình về doanh nghiệp lấy dữ liệu làm trung tâm (data-driven) có kiến trúc hiện đại nhờ kết hợp sức mạnh của cloud (AWS) và khả năng tùy biến cao cấp (kỹ thuật nội bộ)”***

1. **Vai trò của AWS trong từng tầng dữ liệu**

Tầng dữ liệu AWS đóng vai trò

**Ingestion** S3 làm nơi dump log tạm thời; EC2 chạy dịch vụ collect dữ liệu

**Storage** S3 là nơi lưu trữ chính, Iceberg trên S3 là nền tảng bảng dữ liệu

**Serving** Redshift, Druid, DynamoDB phục vụ truy vấn dữ liệu cho BI và ML

<https://aws.amazon.com/vi/solutions/case-studies/netflix-case-study/>

<https://netflixtechblog.com/cloud-efficiency-at-netflix-f2a142955f83>

<https://netflixtechblog.com/cloud-efficiency-at-netflix-f2a142955f83>