**Cùng khám phá hệ thống! Học máy trên AWS**

Bởi Luca Mezzalira, Federica Ciuffo, Vittorio Denti, và Zamira Jaupaj vào ngày 30 tháng 5 năm 2024 trên [Amazon Machine Learning](https://aws.amazon.com/blogs/architecture/category/artificial-intelligence/amazon-machine-learning/), [Amazon SageMaker](https://aws.amazon.com/blogs/architecture/category/artificial-intelligence/sagemaker/), [AWS Inferentia](https://aws.amazon.com/blogs/architecture/category/artificial-intelligence/aws-inferentia/), [AWS Trainium](https://aws.amazon.com/blogs/architecture/category/artificial-intelligence/aws-trainium/) [Permalink](https://aws.amazon.com/blogs/architecture/lets-architect-learn-about-machine-learning-on-aws/) [Share](https://aws.amazon.com/vi/blogs/architecture/lets-architect-learn-about-machine-learning-on-aws/)

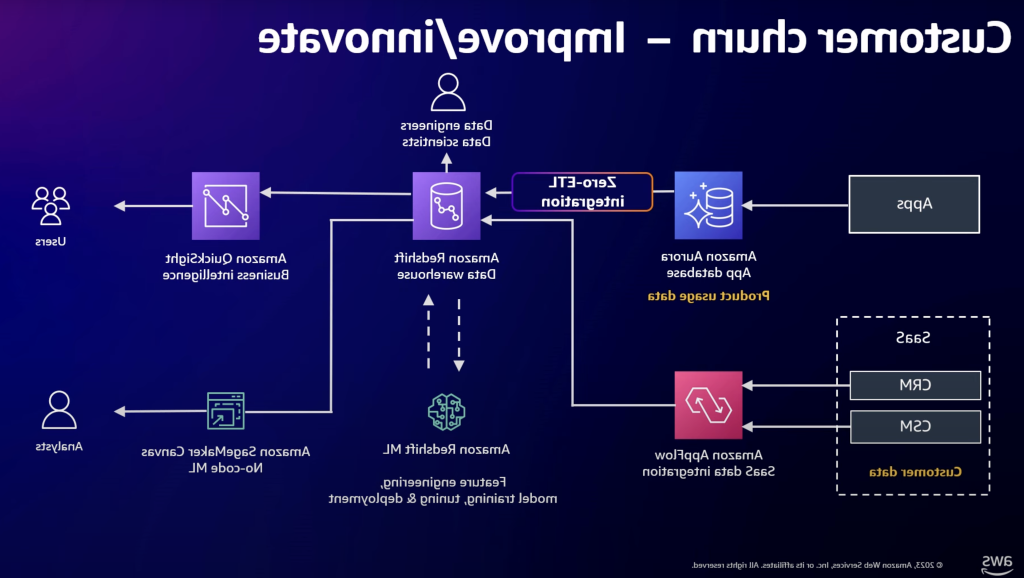
Một cách tiếp cận dựa trên dữ liệu giúp các doanh nghiệp đưa ra quyết định một cách có cơ sở, dựa trên các dự đoán và dự báo chính xác — từ đó nâng cao hiệu quả vận hành và tối ưu hóa nguồn lực. Hệ thống học máy (Machine Learning – ML) có năng lực đặc biệt là khả năng học hỏi và thích nghi liên tục, nhờ tiếp xúc với ngày càng nhiều dữ liệu, hiệu suất của hệ thống được cải thiện theo thời gian. Khả năng tự học này cho phép các tổ chức luôn đi trước xu thế, phản ứng linh hoạt trước những biến động của thị trường và sự thay đổi trong hành vi khách hàng — qua đó thúc đẩy đổi mới sáng tạo và nâng cao năng lực cạnh tranh.

Bằng cách tận dụng sức mạnh của học máy trên nền tảng AWS, các doanh nghiệp có thể khai thác được những lợi ích thiết thực — giúp nâng cao hiệu quả vận hành, cải thiện chất lượng ra quyết định và thúc đẩy tăng trưởng bền vững

[**AWS re:Invent 2023 – Zero to machine learning: Jump-start your data-driven journey**](https://www.youtube.com/watch?v=-CSrOKo8Qgs)

Trong phiên thảo luận này, bạn sẽ thấy cách các tổ chức có nguồn lực hạn chế (về ngân sách, kỹ năng, hoặc thời gian) có thể khởi động hành trình dữ liệu của mình một cách hiệu quả, thông qua phân tích dữ liệu nâng cao và năng lực học máy (ML). Tìm hiểu các bài tập [AWS Working Backwards](https://aws.amazon.com/blogs/opensource/working-backwards-the-story-behind-the-aws-cloud-development-kit/) nhằm thúc đẩy các dự án liên quan đến dữ liệu – tập trung giải quyết các bài toán mang lại giá trị kinh doanh cụ thể. Tiếp theo, đào sâu vào các năng lực phân tích và AI/ML trên AWS giúp đơn giản hóa và đẩy nhanh quá trình triển khai data pipeline cũng như chuyển hóa khối lượng công việc học máy thành giá trị kinh doanh. Khám phá các dịch vụ ít code/không code của AWS trong kiến trúc data pipeline hoàn chỉnh.

[Đưa tôi đến video này](https://www.youtube.com/watch?v=-CSrOKo8Qgs)

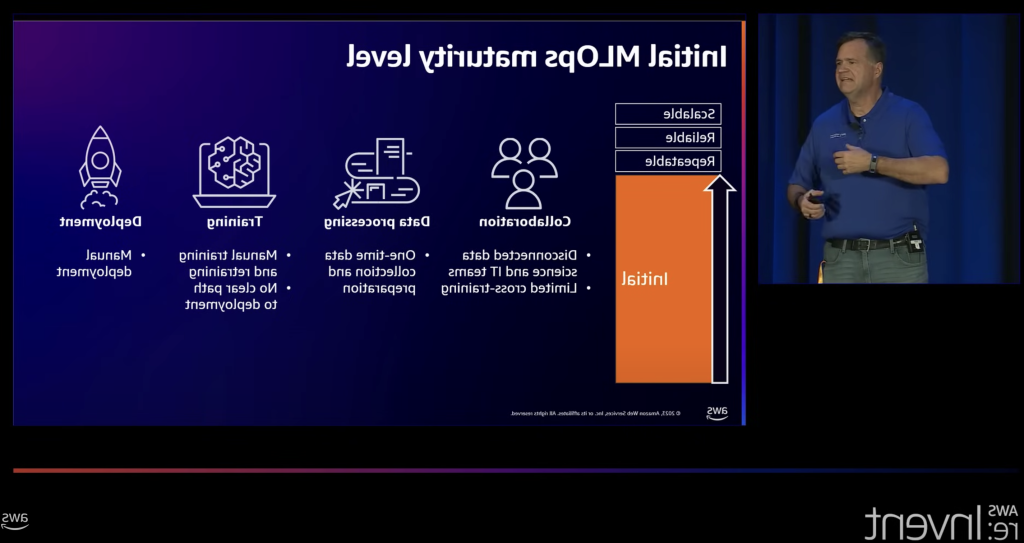


*Hình 1. Hãy xem kiến trúc phân tích tỷ lệ khách hàng rời bỏ sử dụng dịch vụ AWS*

[**Giới thiệu về kỹ thuật MLOps trên AWS**](https://youtu.be/2kzJPhgDkDE?si=vQwL5O-5ag3po49s)

Khi trí tuệ nhân tạo (AI) tiếp tục tạo ra những chuyển biến sâu rộng trong nhiều ngành công nghiệp, thì khả năng vận hành hóa (operationalize) và mở rộng quy mô các mô hình học máy (ML) đang trở thành một thách thức trọng yếu. Phiên thảo luận này giới thiệu về khái niệm MLOps — một lĩnh vực kỹ thuật được xây dựng dựa trên và mở rộng từ các thực tiễn DevOps vốn đã được áp dụng rộng rãi trong phát triển phần mềm. Thông qua việc áp dụng các nguyên tắc của MLOps, các tổ chức có thể tinh gọn quy trình xây dựng, huấn luyện và triển khai mô hình học máy, đảm bảo quản lý vòng đời mô hình một cách hiệu quả và đáng tin cậy. Việc làm chủ MLOps cho phép tổ chức thu hẹp khoảng cách giữa phát triển trí tuệ nhân tạo và vận hành thực tế, từ đó khai phóng toàn bộ tiềm năng từ các sáng kiến học máy.

[Đưa tôi đến video này](https://youtu.be/2kzJPhgDkDE?si=vQwL5O-5ag3po49s)



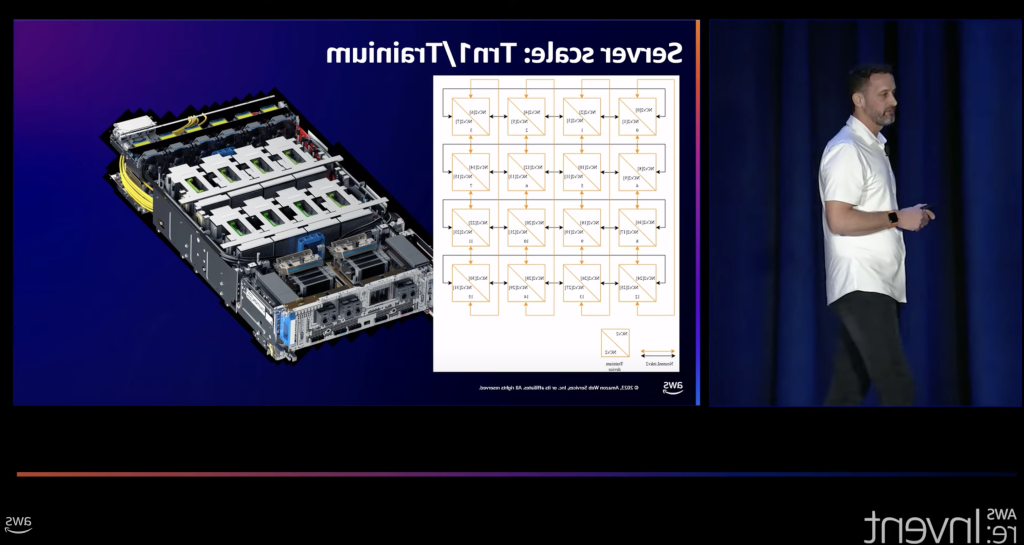
*Hình 2. Mức độ trưởng thành MLOps giúp đánh giá hiện trạng tổ chức và xác định lộ trình nâng cao năng lực.*

[**Đằng sau hậu trường về hạ tầng của AI tạo sinh tại Amazon**](https://www.youtube.com/watch?v=fDk09hms8s8)

Để vận hành hiệu quả các ứng dụng trí tuệ nhân tạo tạo sinh mà vẫn kiểm soát được chi phí, AWS thiết kế và xây dựng các bộ tăng tốc học máy như [AWS Trainium](https://aws.amazon.com/machine-learning/trainium/) và  [AWS Inferentia](https://aws.amazon.com/machine-learning/inferentia/). Phiên thảo luận này sẽ giới thiệu các phần cứng học máy chuyên biệt cho hai giai đoạn chính: huấn luyện mô hình và suy luận mô hình. Đồng thời, trình bày cách Amazon cũng như khách hàng của AWS tận dụng các giải pháp này để tối ưu hóa chi phí và giảm độ trễ.

Bạn sẽ được tìm hiểu thông qua các ví dụ thực tiễn cho thấy tác động của các giải pháp phần cứng này, cùng với phần giải thích về nguyên lý hoạt động của các chip tăng tốc học máy. Các bộ tăng tốc học máy không chỉ hữu ích cho khối lượng công việc trí tuệ nhân tạo tạo sinh mà còn có thể áp dụng vào nhiều tình huống khác như học biểu diễn đặc trưng, hệ thống gợi ý và bất kỳ bài toán nào liên quan đến mô hình mạng nơ-ron sâu.

[Đưa tôi đến video này](https://www.youtube.com/watch?v=fDk09hms8s8)



*Hình 3. Khám phá công nghệ đằng sau các dịch vụ trí tuệ nhân tạo của AWS*

**Cách khách hàng của chúng tôi triển khai học máy trên AWS**

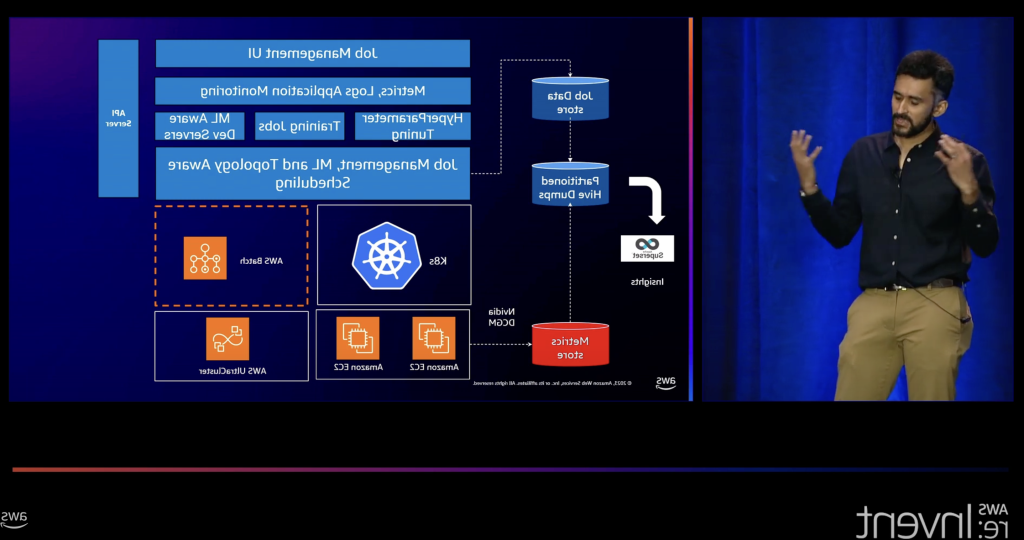
Các tài nguyên dưới đây đi sâu vào hạ tầng học máy được sử dụng để huấn luyện các mô hình lớn tại [Pinterest](https://www.pinterest.com/) và thử nghiệm dưới các khung được chế tạo bởi [Booking.com](https://www.booking.com/).

Trong video của Pinterest, bạn sẽ tìm hiểu chiến lược thiết lập môi trường phát triển học máy, điều phối các bài huấn luyện, tích hợp dữ liệu vào vòng lặp huấn luyện, và tăng tốc độ huấn luyện mô hình. Bên cạnh đó, video còn trình bày lợi ích của việc sử dụng container trong bối cảnh học máy, và cách Pinterest triển khai toàn bộ vòng đời học máy, bao gồm cả huấn luyện mô hình phân tán.

Tài nguyên thứ hai trình bày cách Booking rút ngắn thời gian thử nghiệm bằng cách tận dụng Amazon SageMaker cho các giai đoạn phân tích dữ liệu, huấn luyện mô hình, và thử nghiệm trực tuyến. Kết quả là họ đã giảm đáng kể thời gian phát triển các mô hình xếp hạng, đồng thời tăng tốc độ làm việc của các nhóm khoa học dữ liệu.

[Đưa tôi đến video của Pinterest](https://www.youtube.com/watch?v=wM0iSWAzyhI)

[Đưa tôi đến bài blog của Booking.com](https://aws.amazon.com/blogs/machine-learning/how-booking-com-modernized-its-ml-experimentation-framework-with-amazon-sagemaker/)

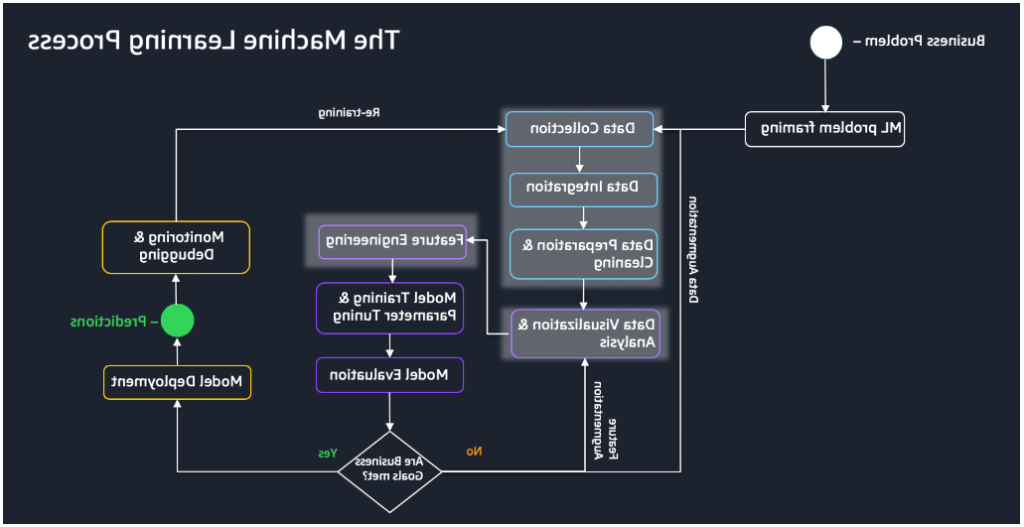


*Hình 4. Khám phá cách Pinterest sử dụng các dịch vụ AWS cho khối lượng các công việc học máy*

[**Ngày trải nghiệm chuyên sâu với SageMaker**](https://catalog.us-east-1.prod.workshops.aws/workshops/63069e26-921c-4ce1-9cc7-dd882ff62575/en-US)

Ngày trải nghiệm chuyên sâu với SageMakerhelps của Amazon là chương trình dành cho khách hàng và đối tác, cung cấp cái nhìn toàn diện về cách xây dựng các bài toán học máy thực tế. Từ kỹ thuật xây dựng đặc trưng cho đến việc tìm hiểu các thuật toán tích hợp sẵn, chương trình tập trung vào quy trình huấn luyện, hiệu chỉnh, và triển khai mô hình học máy trong môi trường gần giống sản xuất, hội thảo này hướng dẫn bạn mang theo mô hình của chính mình, và thực hiện quá trình chuyển đổi hệ thống từ môi trường cơ sở lên nền tảng Amazon SageMaker. Ngoài ra, chương trình còn giới thiệu các khái niệm nâng cao như gỡ lỗi mô hình, giám sát mô hình, tự động hóa học máy.

[Đưa tôi đến buổi hội thảo](https://catalog.us-east-1.prod.workshops.aws/workshops/63069e26-921c-4ce1-9cc7-dd882ff62575/en-US)



*Hình 5. Huấn luyện, hiệu chỉnh và triển khai khối lượng công việc của bạn với Amazon SageMaker*

**Hẹn gặp lại ở lần tới!**

Cảm ơn bạn đã theo dõi! Qua bài viết này, chúng tôi đã giới thiệu đến bạn những khả năng tiềm năng khi ứng dụng các dịch vụ học máy của AWS. Trong bài viết tiếp theo, chúng tôi sẽ thảo luận về di chuyển hệ thống lên đám mây.

Để xem lại các bài viết trước hoặc khám phá toàn bộ chuỗi bài, hãy truy cập trang [Let’s Architect!](https://aws.amazon.com/blogs/architecture/tag/lets-architect/) .

TAGS: [Amazon Sagemaker](https://aws.amazon.com/blogs/architecture/tag/amazon-sagemaker/), [artificial intelligence](https://aws.amazon.com/blogs/architecture/tag/artificial-intelligence/), [data analytics](https://aws.amazon.com/blogs/architecture/tag/data-analytics/), [Let's Architect](https://aws.amazon.com/blogs/architecture/tag/lets-architect/), [machine learning](https://aws.amazon.com/blogs/architecture/tag/machine-learning/), [Software as a Service](https://aws.amazon.com/blogs/architecture/tag/software-as-a-service/)



**Luca Mezzalira**

Luca hiện là Kiến trúc sư Giải pháp Cấp cao làm việc tại London. Ông là tác giả của nhiều cuốn sách và là diễn giả quốc tế. Chuyên môn của ông tập trung chủ yếu trong lĩnh vực kiến trúc giải pháp. Luca đã được ghi nhận nhờ những đóng góp mang tính đột phá trong việc mở rộng quy mô kiến trúc giao diện người dùng thông qua mô hình kiến trúc phân tách giao diện người dùng thành các mô-đun độc lập — từ việc nâng cao hiệu suất quy trình làm việc cho đến đảm bảo chất lượng sản phẩm đầu ra.



**Federica Ciuffo**

Federica là Kiến trúc sư Giải pháp tại Amazon Web Services (AWS). Cô chuyên về các dịch vụ container và đặc biệt đam mê xây dựng hạ tầng bằng mã.Ngoài công việc, Federica yêu thích đọc sách, vẽ tranh và dành thời gian bên bạn bè — đặc biệt là tại các nhà hàng, nơi cô có thể khám phá những món ăn mới từ nhiều nền ẩm thực khác nhau.



**Vittorio Denti**

Vittorio Denti là Kỹ sư Học máy tại Amazon trụ sở London.Sau khi hoàn thành chương trình Thạc sĩ Khoa học Máy tính và Kỹ thuật tại đại học Politecnico di Milano và Viện Công nghệ Hoàng gia KTH ở Stockholm, anh đã gia nhập AWS. Vittorio có nền tảng chuyên môn về hệ thống phân tán và học máy. Anh đặc biệt đam mê kỹ thuật phần mềm và các đổi mới tiên phong trong lĩnh vực khoa học học máy.



**Zamira Jaupaj**

Zamira là Kiến trúc sư Giải pháp Doanh nghiệp làm việc tại Hà Lan. Cô là một chuyên gia công nghệ đầy đam mê, với hơn 10 năm kinh nghiệm làm việc trong môi trường đa quốc gia, chuyên thiết kế và triển khai các giải pháp trọng yếu và phức tạp sử dụng container, kiến trúc serverless và phân tích dữ liệu — phục vụ cả doanh nghiệp vừa và nhỏ lẫn các tập đoàn lớn.