

ĐẠI HỌC QUỐC GIA TP. HCM
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA HỆ THÔNG THÔNG TIN



ĐỒ ÁN MÔN HỌC
ĐIỆN TOÁN ĐÁM MÂY
ĐỀ TÀI: XÂY DỰNG HỆ THỐNG ĐỀ NGHỊ PHIM
TRÊN NỀN TẢNG ĐIỆN TOÁN ĐÁM MÂY

Giảng viên hướng dẫn: ThS. Hà Lê Hoài Trung

Lớp: IS402.O11

Nhóm sinh viên thực hiện

Nguyễn Đạt	MSSV: 20520434
Trần Thị Ngọc Ánh	MSSV: 20521083
Nguyễn Minh Cường	MSSV: 20520422
Tôn Nữ Thảo Nhi	MSSV: 20521718

TP. HỒ CHÍ MINH 12/2023

LỜI CẢM ƠN

Trên thực tế không có sự thành công nào mà không gắn liền với những sự hỗ trợ, giúp đỡ dù ít hay nhiều, dù trực tiếp hay gián tiếp của người khác. Với lòng biết ơn sâu sắc nhất, nhóm chúng em xin gửi lời cảm ơn đến tập thể quý Thầy Cô Trường Đại học Công nghệ thông tin – ĐHQG TP.HCM và quý Thầy Cô khoa Hệ thống thông tin đã giúp cho nhóm chúng em có những kiến thức cơ bản làm nền tảng để thực hiện đề tài này. Đặc biệt nhóm chúng em xin gửi lời cảm ơn chân thành tới thầy Hà Lê Hoài Trung – giảng viên lý thuyết môn Điện toán đám mây đã tận tình giúp đỡ, trực tiếp chỉ bảo, hướng dẫn nhóm trong suốt quá trình làm đồ án môn học. Nhờ đó, chúng em đã tiếp thu được nhiều kiến thức bổ ích trong việc vận dụng cũng như kỹ năng làm đồ án. Nếu không có những lời hướng dẫn, dạy bảo của thầy thì nhóm chúng em nghĩ đồ án này của nhóm rất khó có thể hoàn thiện được.

Ngoài ra, để đồ án được hoàn thành thì không thể nào không cảm ơn những người đã làm ra nó, cảm ơn các thành viên trong nhóm đã chăm chỉ và chịu khó hoàn thành nhiệm vụ đúng tiến độ. Dựa trên những kiến thức được thầy cung cấp trên trường kết hợp với việc tự tìm hiểu về những kiến thức mới, nhóm đã cố gắng thực hiện đồ án một cách tốt nhất. Từ đó, nhóm chúng em vận dụng tối đa những gì đã thu thập được để hoàn thành một báo cáo đồ án tốt nhất. Tuy nhiên, trong quá trình thực hiện, nhóm chúng em không tránh khỏi những thiếu sót. Chính vì vậy, nhóm chúng em rất mong nhận được những sự góp ý từ phía thầy nhằm hoàn thiện những kiến thức mà nhóm em đã học tập và là hành trang để nhóm chúng em thực hiện tiếp các đề tài khác trong tương lai.

Sau cùng, chúng em xin kính chúc thầy thật dồi dào sức khỏe, niềm tin để tiếp tục thực hiện sứ mệnh cao đẹp là truyền đạt kiến thức cho các bạn sinh viên.

TP. Hồ Chí Minh, tháng 12, năm 2023

Nhóm thực hiện

NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

MỤC LỤC

LỜI CẢM ƠN	2
NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN	3
MỤC LỤC.....	4
DANH MỤC BẢNG.....	6
DANH MỤC HÌNH ẢNH	6
CHƯƠNG 1. TỔNG QUAN ĐỀ TÀI	10
1.1. Lý do chọn đề tài	10
1.2. Mô tả bài toán.....	11
1.3. Mô tả dataset	12
1.4. Quy trình thực hiện	14
CHƯƠNG 2. GIỚI THIỆU VỀ HỆ THỐNG AZURE	19
2.1. Giới thiệu Microsoft Azure	19
2.2. Azure Blob Storage	21
2.3. Azure Data Factory	23
2.4. Azure Databricks.....	24
2.5. Azure Machine Learning Service.....	25
2.6. Azure Cosmos DB.....	26
2.7. Azure Container Instances	27
2.8. Azure Kubernetes Service.....	28
2.9. PowerApps	30
CHƯƠNG 3. QUY TRÌNH RA QUYẾT ĐỊNH	32
3.1. Phương pháp đề xuất phim.....	32

3.2.	Phương pháp đánh giá.....	36
3.3.	Thuật toán Alternating Least Square.....	36
3.4.	Thuật toán Surprise Singular Value Decomposition.....	38
3.5.	Bảng so sánh kết quả.....	40
3.6.	Trực quan hóa dữ liệu	41
CHƯƠNG 4. CÁC BƯỚC TRIỂN KHAI.....		42
4.1.	Triển khai môi trường Azure	42
4.2.	Triển khai môi trường Azure Storage	45
4.3.	Triển khai môi trường Azure Databricks	49
4.4.	Triển khai môi trường Azure Data Factory.....	53
4.5.	Triển khai môi trường Azure Cosmos DB	61
4.6.	Triển khai môi trường Azure Machine Learning Service	63
4.7.	Triển khai môi trường Azure Kubernetes Service	64
4.8.	Triển khai ứng dụng trên Power Apps	67
4.9.	Triển khai toàn bộ hệ thống	70
CHƯƠNG 5. KẾT LUẬN		72
5.1.	Kết quả đạt được	72
5.2.	Hạn chế.....	72
5.3.	Hướng phát triển	72
BẢNG PHÂN CÔNG CÔNG VIỆC.....		74
TÀI LIỆU THAM KHẢO.....		75

DANH MỤC BẢNG

Bảng 1.1. Table users	13
Bảng 1.2. Table movies.....	13
Bảng 1.3. Table history_watch.....	14
Bảng 2.1. Bảng so sánh các loại Blobs	23
Bảng 3.1. Bảng so sánh kết quả	40
Bảng 1. Bảng phân công công việc	74

DANH MỤC HÌNH ẢNH

Hình 1.1. Quy trình thực hiện đề tài.....	14
Hình 1.2. Quy trình nhóm thực hiện	15
Hình 2.1. Microsoft Azure	19
Hình 2.2. Các dịch vụ Azure cung cấp.....	20
Hình 2.3. Microsoft Azure Blob Storage	21
Hình 2.4. Minh họa mối quan hệ giữa các resources	21
Hình 2.5. Các loại Blob.....	22
Hình 2.6. Azure Data Factory	23
Hình 2.7. Azure Databricks.....	24
Hình 2.8. Azure Machine Learning Service.....	25
Hình 2.9. Azure Cosmos DB.....	26
Hình 2.10. Azure Container Instances	27
Hình 2.11. Azure Kubernetes Service.....	28
Hình 2.12. Power Apps	30
Hình 2.13. Ứng dụng có thể chạy trên trình duyệt và di động.....	30
Hình 3.1. Gợi ý phim dựa trên sở thích của người dùng.....	32
Hình 3.2. Item-based collaborative filtering	33
Hình 3.3. Ví dụ về CF	34

Hình 3.4. Mô phỏng quá trình gợi ý phim theo phương pháp item-item Collaborative Filtering	35
Hình 3.5. Phương pháp lắp trong ALS.....	37
Hình 3.6. Mô phỏng dữ liệu đánh giá của khách hàng cho từng bộ phim	37
Hình 3.7. Kết quả khi sử dụng ALS.....	38
Hình 3.8. Thuật toán ALS	41
Hình 3.9. Thuật toán SVD.....	41
Hình 4.1. Tìm kiếm Service Resource Groups	42
Hình 4.2. Tạo Resource Groups	43
Hình 4.3. Điền các thông tin cho Resource Groups.....	43
Hình 4.4. Kiểm tra thông tin trước khi khởi tạo.....	44
Hình 4.5. Kết quả khởi tạo Resouce Groups.....	44
Hình 4.6. Tìm kiếm Storage Accounts.....	45
Hình 4.7. Tạo mới Storage Accounts	45
Hình 4.8. Khởi tạo các thông tin cơ bản của Storage tại Base.....	46
Hình 4.9. Khởi tạo các thông tin cơ bản của Storage tại Advanced	46
Hình 4.10. Khởi tạo các thông tin cơ bản của Storage tại Networking	47
Hình 4.11. Khởi tạo các thông tin cơ bản của Storage tại Data protection.....	47
Hình 4.12. Khởi tạo các thông tin cơ bản của Storage tại Encyption and Tags	48
Hình 4.13. Review lại thông tin Storage Acount trước khi Create tại tag Review	48
Hình 4.14. Khởi tạo Storage accounts thành công.....	49
Hình 4.15. Tìm kiếm Services Azure Databricks	49
Hình 4.16. Tạo mới Azure Databricks	50
Hình 4.17. Thiết lập thông tin cơ bản tại Basics	50
Hình 4.18. Thiết lập Netwworking để khởi tạo Databricks	51
Hình 4.19. Thiết lập các thông tin Ancrytion and Tags.....	51
Hình 4.20. Đợi kết quả Deloy	52
Hình 4.21. Màn hình sau khi khởi tạo Azure Databricks.....	52
Hình 4.22. Tiến hành tạo mới Cluster với các thông số.....	53

Hình 4.23. Tìm kiếm Data Factory	53
Hình 4.24. Cấu hình thông tin cơ bản tại Base	54
Hình 4.25. Cấu hình liên kết tới Github.....	54
Hình 4.26. Cấu hình networking	55
Hình 4.27. Cấu hình advanced	55
Hình 4.28. Cấu hình Tags	56
Hình 4.29. Review lại thông tin và khởi tạo	56
Hình 4.30. Kết quả khởi tạo Data factories.....	57
Hình 4.31. Khởi tạo link service tới SQL	57
Hình 4.32. Khởi tạo link service tới API TMDB	58
Hình 4.33. Khởi tạo link service tới Blob	58
Hình 4.34. Lấy thông tin dữ liệu tổng số trang	59
Hình 4.35. Biến đổi kết quả trả về dữ liệu để get total page.....	59
Hình 4.36. Vòng lặp Foreach duyệt qua lần lượt các trang	60
Hình 4.37. Cấu hình Source copy from TMDB	60
Hình 4.38. Cấu hình Sink to Blob Storage.....	61
Hình 4.39. Khởi tạo Resource Azure Cosmos DB for NoSQL	61
Hình 4.40. Xem các Database đã tạo	62
Hình 4.41. Thông tin về Database đã khởi tạo.....	62
Hình 4.42. Thông tin về container trong DataExplorer	63
Hình 4.43. Khởi tạo Resource Azure ML	63
Hình 4.44. Thông tổng quát Resource Azure ML đã tạo	64
Hình 4.45. Giao diện Azure ML studio.....	64
Hình 4.46. Khởi tạo AksCompute trong Azure ML studio.....	65
Hình 4.47. Cài đặt các thông số cho AksCompute	65
Hình 4.48. Danh sách các Endpoint trong Azure ML studio	66
Hình 4.49. Thông tin chi tiết AksCompute - 1.....	66
Hình 4.50. Thông tin chi tiết AksCompute - 2.....	67
Hình 4.51. Màn hình đăng nhập vào hệ thống	67

Hình 4.52. Màn hình đăng ký tài khoản.....	68
Hình 4.53. Màn hình tìm kiếm bộ phim và đánh giá	68
Hình 4.54. Màn hình xem thông tin chi tiết phim và phim recommendation.....	69
Hình 4.55. Màn hình xem trailer Phim	69
Hình 4.56. Màn hình đánh giá phim	70
Hình 4.57. Màn hình xuất thông báo.....	70
Hình 4.58. Cài đặt file JAR trong Libraries của Cluster.....	71

CHƯƠNG 1. TỔNG QUAN ĐỀ TÀI

1.1. Lý do chọn đề tài

Đề tài "Xây dựng hệ thống đề nghị phim trên nền tảng điện toán đám mây" là sự kết hợp giữa công nghệ trí tuệ nhân tạo (AI) và điện toán đám mây, hai trong số những lĩnh vực công nghệ hàng đầu và phát triển nhanh chóng hiện nay. Mục tiêu của dự án là phát triển một hệ thống có khả năng phân tích và gợi ý phim dựa trên những bộ phim trước đó có độ tương đồng tương tự, qua đó mang đến trải nghiệm cá nhân hóa và độc đáo cho mỗi người dùng.

Nhóm chúng em đã chọn Python và cụ thể hóa sử dụng PySpark trên Databricks làm ngôn ngữ lập trình chính. PySpark, một giao diện của Apache Spark cho Python, được chọn vì khả năng xử lý dữ liệu lớn một cách hiệu quả, đặc biệt trong môi trường phân tán như điện toán đám mây. Việc tích hợp PySpark với Databricks giúp chúng tôi tận dụng lợi thế của môi trường điện toán đám mây như khả năng mở rộng và quản lý tài nguyên một cách linh hoạt, đồng thời hỗ trợ mạnh mẽ cho các tác vụ phân tích dữ liệu phức tạp và xây dựng mô hình học máy.

Sự lựa chọn triển khai hệ thống trên nền tảng điện toán đám mây Azure mang lại những lợi ích đáng kể. Sử dụng Azure Blob Storage cho phép chúng tôi lưu trữ và quản lý lượng lớn dữ liệu đánh giá phim một cách hiệu quả, trong khi Azure Kubernetes Service đảm bảo triển khai mô hình học máy một cách ổn định và linh hoạt.

Mục tiêu chính của dự án không chỉ là áp dụng và phát triển kiến thức về công nghệ thông tin, mà còn góp phần tạo ra giá trị mới trong lĩnh vực giải trí số. Hệ thống đề xuất phim cá nhân hóa sẽ giúp nâng cao trải nghiệm người dùng và tạo ra giá trị kinh doanh cho các nền tảng phát sóng trực tuyến. Dự án này cũng mở đường cho các nghiên cứu và phát triển tiếp theo trong lĩnh vực AI và điện toán đám mây, hướng tới việc đáp ứng nhu cầu ngày càng cao và đa dạng của thị trường giải trí hiện đại.

Thông qua việc thực hiện đề tài này, nhóm không chỉ có cơ hội thử thách và phát triển chuyên môn mà còn hứa hẹn tạo ra tác động tích cực đối với ngành công nghiệp giải trí và

công nghệ. Việc cung cấp các gợi ý phim chính xác và phù hợp với từng cá nhân không chỉ giúp tăng cường sự hài lòng và trung thành của người dùng mà còn góp phần quan trọng trong việc phân tích xu hướng và sở thích của khách hàng. Điều này hỗ trợ quyết định kinh doanh và chiến lược nội dung, đồng thời mở ra hướng phát triển mới cho các nền tảng truyền thông và giải trí.

Dự án "Xây dựng hệ thống đề nghị phim trên nền tảng điện toán đám mây" không chỉ đánh dấu một bước tiến trong sự nghiệp của chúng tôi mà còn là minh chứng cho sự kết hợp hiệu quả giữa công nghệ thông tin và giải trí. Sự kết hợp này mở ra cánh cửa cho những phát minh, ứng dụng mới và tạo dựng một nền tảng vững chắc cho những nghiên cứu tiếp theo trong lĩnh vực trí tuệ nhân tạo và điện toán đám mây. Chúng tôi tin tưởng rằng, với sự đầu tư và phát triển liên tục, hệ thống này sẽ không chỉ đem lại giá trị thực tiễn mà còn mở ra những cơ hội mới trong việc tối ưu hóa và cá nhân hóa trải nghiệm giải trí số.

Tóm lại, dự án này không chỉ là một thách thức chuyên môn mà còn là một cơ hội để chúng tôi đóng góp vào sự phát triển của ngành công nghiệp giải trí số và công nghệ thông tin. Với sự sáng tạo, đổi mới và áp dụng công nghệ tiên tiến, chúng tôi hy vọng sẽ tạo ra một hệ thống đề nghị phim không chỉ mạnh mẽ và thông minh mà còn thân thiện và dễ tiếp cận với người dùng cuối, góp phần làm phong phú thêm trải nghiệm giải trí số trong thời đại công nghệ 4.0.

1.2. Mô tả bài toán

Trong thế giới số hóa và trực tuyến ngày nay, việc tạo ra một hệ thống đề xuất phim cá nhân hóa chính xác và hiệu quả đã trở thành một bài toán cấp thiết và thách thức. Hệ thống đề xuất phim sử dụng phương pháp lọc cộng tác (Collaborative Filtering - CF) nhằm giải quyết bài toán này bằng cách phân tích và học hỏi từ dữ liệu hành vi và sở thích của người dùng.

Bài toán cụ thể là làm thế nào để hệ thống có thể xác định và gợi ý các bộ phim mà người dùng có khả năng quan tâm và thích thú. Điều này đòi hỏi hệ thống phải có khả năng:

- **Thu Thập và Phân Tích Dữ Liệu:** Hệ thống cần thu thập dữ liệu từ hoạt động xem phim của người dùng, bao gồm các bộ phim họ đã xem, thời gian xem, đánh giá phim, và các tương tác khác trên nền tảng.
- **Xác Định Mối Quan Hệ Giữa Người Dùng:** Sử dụng dữ liệu đã thu thập, hệ thống cần phân tích và xác định mối quan hệ giữa các người dùng dựa trên sở thích phim tương tự. Điều này giúp hệ thống hiểu được nhóm người dùng nào có sở thích phim giống nhau.
- **Dựa Vào Mối Quan Hệ Để Đề Xuất:** Dựa trên thông tin từ nhóm người dùng có sở thích tương tự, hệ thống sẽ gợi ý những bộ phim mà người dùng có thể chưa biết nhưng có khả năng sẽ thích.
- **Tinh Chỉnh và Cập Nhật Liên Tục:** Hệ thống cần được tinh chỉnh và cập nhật liên tục dựa trên phản hồi của người dùng để cải thiện độ chính xác của các gợi ý.

Đối mặt với thách thức này, mục tiêu của hệ thống là cung cấp trải nghiệm xem phim cá nhân hóa, tăng cường sự hài lòng và gắn kết của người dùng với nền tảng, đồng thời tạo ra một phương pháp tiếp cận thông minh và tự động hóa trong việc giới thiệu nội dung giải trí.

1.3. Mô tả dataset

Đây là bộ dữ liệu được lấy trên trang TMDB là viết tắt của “The Movie Database”, một cơ sở dữ liệu trực tuyến lớn chứa thông tin về phim, bao gồm thông tin về diễn viên, đạo diễn, biên kịch, hình ảnh, bài nhạc, và nhiều thông tin khác liên quan đến ngành công nghiệp điện ảnh. TMDB cung cấp các API miễn phí có thể sử dụng để truy cập các dữ liệu trên. Bộ dữ liệu ‘movies’ được trích xuất từ API lấy các bộ phim phổ biến (<https://api.themoviedb.org/3/person/popular>). Các tập dữ liệu về ‘users’ và ‘history_watch’ được tạo ngẫu nhiên dựa trên tập dữ liệu ‘movie’ và các thư viện hỗ trợ tạo dữ liệu mẫu trong Python.

STT	Tên thuộc tính	Kiểu dữ liệu	Ý nghĩa
1	user_id	integer	Mã người dùng
2	fullname	string	Họ tên người dùng
3	gender	string	Giới tính người dùng (male và female)
4	birth_year	integer	Năm sinh của người dùng
5	address	string	Địa chỉ người dùng
6	username	string	Tên đăng nhập của người dùng
7	password	string	Mật khẩu đăng nhập của người dùng

Bảng 1.1. Table users

Tổng số người dùng: 1999

STT	Tên thuộc tính	Kiểu dữ liệu	Ý nghĩa
1	id	integer	Mã bộ phim
2	name	string	Tên bộ phim
3	category	string	Thể loại bộ phim
4	year	integer	Năm chiếu của bộ phim
5	country	string	Quốc gia phát hành bộ phim
6	rating_avg	float	Đánh giá trung bình của bộ phim

Bảng 1.2. Table movies

Tổng số lượng đánh giá: 44,878

STT	Tên thuộc tính	Kiểu dữ liệu	Ý nghĩa
1	user_id	integer	Mã người dùng
2	movie_id	integer	Mã bộ phim
3	like	boolean	Người dùng có thích bộ phim đó hay không
4	date_watch	date	Ngày người dùng xem bộ phim đó

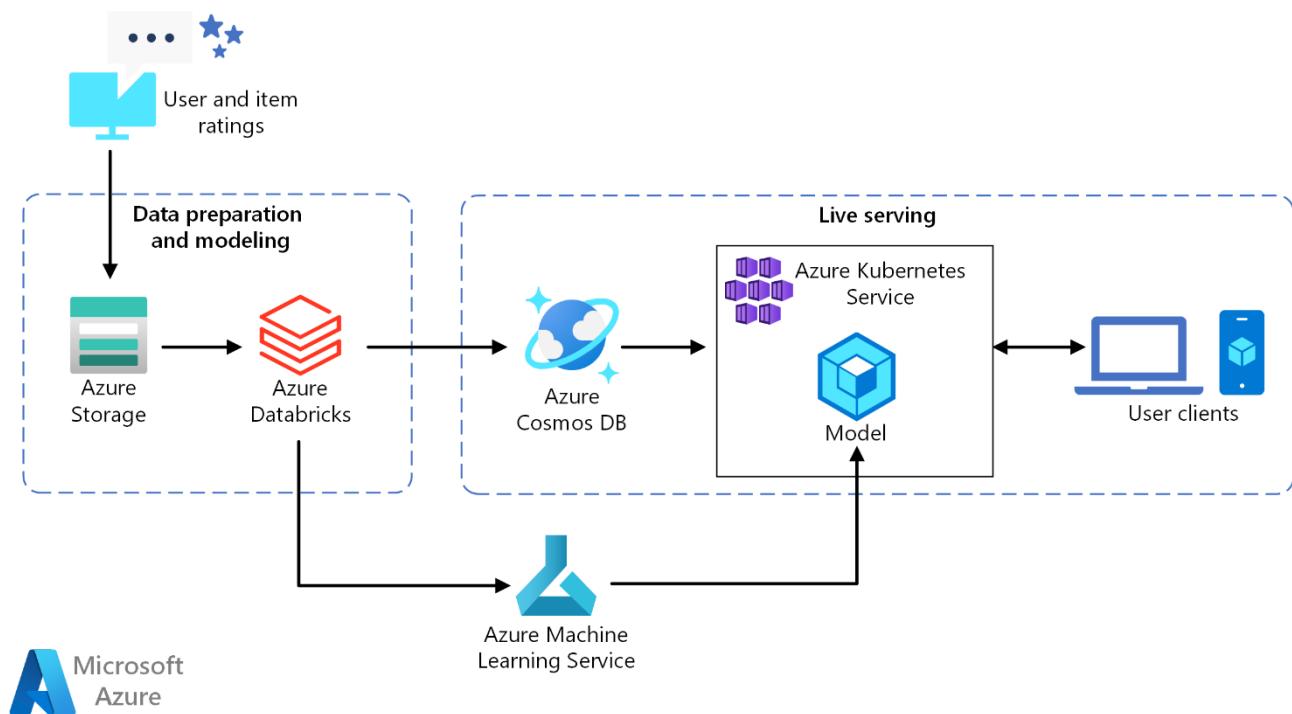
5	rating	float	Đánh giá của người dùng đối với bộ phim
---	--------	-------	---

Bảng 1.3. Table history_watch

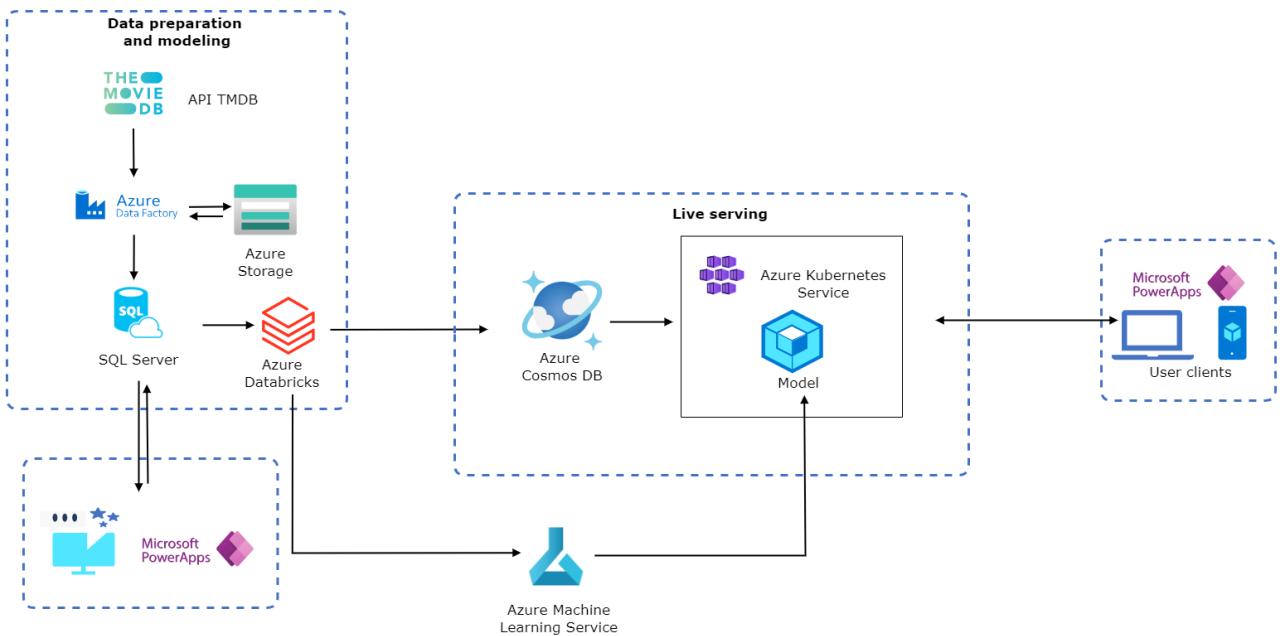
Tổng số lượng bộ phim: 547 bộ phim

1.4. Quy trình thực hiện

Một hệ thống chuẩn hệ hỗ trợ ra quyết định được mô tả như hình dưới đây:



Hình 1.1. Quy trình thực hiện để tài



Hình 1.2. Quy trình nhóm thực hiện

- Quy trình thực hiện bài toán hệ thống đề xuất thời gian thực theo phương pháp Collaborative Filtering với hai thuật toán Singular Value Decomposition – SVD và Alternating Least Squares – ALS trên môi trường Azure được mô tả cơ bản như sau:

- 1. Thu thập và giả lập dữ liệu đầu vào:** Bắt đầu với việc thu thập dữ liệu đánh giá từ người dùng và thông tin sản phẩm. Nhóm thực hiện sẽ tạo giả lập dữ liệu đầu vào bằng cách nghiên cứu và phân tích các dataset khác, tích hợp thông tin và các chiều thông tin khác nhau để phục vụ việc train model và đánh giá các giải thuật.
- 2. Thu thập dữ liệu master từ TMDB:** Trong bài toán này, dữ liệu thông tin về các bộ phim là dữ liệu master mà nhóm cần thu thập và chuẩn bị, nhóm sử dụng Azure Data Factory để gọi API tới TMDB lấy dữ liệu master cho các bộ phim, điều này phục vụ cho việc sau này người dùng có thể tiến hành vào chọn thông tin của các bộ phim.
- 3. Tương tác người dùng qua Power Apps:** Nhóm sử dụng Power Apps để tạo một ứng dụng giao diện thân thiện với người dùng cho người dùng tiến hành đánh giá, bình luận và chọn lọc các bộ phim.

- 4. Lưu trữ dữ liệu và phân bổ dữ liệu:** Các thông tin master data không thể thay đổi thì sẽ được lưu trữ trên SQL Server và các thông tin về dữ liệu người dùng đánh giá phim cũng như các thông tin cần sửa đổi sẽ được lưu trữ trên SharePoint để tiện cho việc quản lý và cuối cùng tất cả dữ liệu được lưu trữ vào SQL Server và cả trong Azure Storage để tiện quản lý và theo dõi và tuỳ mục đích mà chúng ta có thể gọi và xử lý sau này.
- 5. Tiền xử lý dữ liệu bằng Azure Databricks:** Nhóm sử dụng Azure Databricks để tiền xử lý dữ liệu từ SQL Server. Quá trình này bao gồm việc làm sạch dữ liệu, chọn lọc các thuộc tính, và chuyển đổi dữ liệu thành định dạng phù hợp cho việc huấn luyện mô hình.
- 6. Huấn luyện mô hình bằng Azure Databricks:** Nhóm tiếp tục sử dụng Azure Databricks để huấn luyện các mô hình học máy. Trong trường hợp này, SVD và ALS là hai thuật toán được sử dụng để phát hiện các yếu tố tiềm ẩn dựa trên dữ liệu đánh giá của người dùng đối với các bộ phim. Việc huấn luyện mô hình yêu cầu việc tối ưu hóa các tham số để đạt được độ chính xác dự đoán cao nhất có thể.
- 7. Đánh giá và Tinh chỉnh Mô hình:** Sau khi huấn luyện, mô hình được đánh giá để xác định hiệu suất của nó. Nhóm sử dụng hai độ đo là RMSE và MAE có thể được sử dụng để đánh giá. Tinh chỉnh mô hình có thể bao gồm việc điều chỉnh các hyperparameters hoặc thử nghiệm với các kỹ thuật khác nhau để tối ưu hóa model.
- 8. Lưu trữ và quản lý mô hình với Azure Cosmos DB và Machine Learning Azure:** Mô hình đã được tinh chỉnh sau đó được lưu trữ trong Azure Cosmos DB và sử dụng Machine Learning Azure để hỗ trợ deploy, cho phép truy cập nhanh chóng và dễ dàng khi cần phục vụ đề xuất.
- 9. Triển khai Mô hình với Azure Kubernetes Service (AKS):** Mô hình được triển khai sử dụng Azure Kubernetes Service (AKS) để phục vụ đề xuất thời gian thực. AKS cung cấp khả năng tự động mở rộng và quản lý các dịch vụ container hóa.

10. Phục vụ Đề xuất: Khi người dùng tương tác với hệ thống, AKS giao tiếp với mô hình trong Cosmos DB để cung cấp đề xuất dựa trên hành vi và sở thích của họ.

11. Theo dõi và Cập nhật Mô hình: Hệ thống cần được giám sát liên tục để đảm bảo rằng nó hoạt động hiệu quả. Mô hình có thể cần được cập nhật thường xuyên để phản ánh sự thay đổi trong hành vi người dùng hoặc để cải thiện chất lượng đề xuất.

Với quy trình trên, nhấn mạnh vào việc lưu trữ dữ liệu một cách linh hoạt và an toàn, cũng như việc tối ưu hóa quản lý và xử lý dữ liệu thông qua việc sử dụng các dịch vụ khác nhau của Azure. Điều này cũng đảm bảo rằng hệ thống có khả năng thu thập, lưu trữ, và xử lý dữ liệu một cách hiệu quả, đồng thời duy trì sự linh hoạt cần thiết để thích ứng với sự thay đổi trong yêu cầu của người dùng. Ngoài ra hệ thống cũng cần đảm bảo được các yêu cầu:

- Tích hợp dữ liệu và đảm bảo bảo mật:** Tích hợp dữ liệu từ các nguồn khác nhau (SQL Server, SharePoint, Azure Storage) và đảm bảo rằng dữ liệu được xử lý và lưu trữ một cách an toàn, tuân thủ các chuẩn mực bảo mật và quy định về dữ liệu.
- Phân tích và tối ưu hóa hệ thống:** Tiến hành phân tích liên tục để đánh giá hiệu suất của hệ thống. Sử dụng những phân tích này để tối ưu hóa quy trình xử lý dữ liệu, huấn luyện mô hình, và phục vụ đề xuất.
- Mở rộng và quy mô hóa:** Xem xét việc mở rộng hệ thống và tăng cường các nguồn lực khi cần thiết để đáp ứng với sự tăng trưởng về số lượng người dùng và khối lượng dữ liệu.
- Cập nhật và bảo trì hệ thống:** Đảm bảo rằng hệ thống luôn được cập nhật với các công nghệ mới nhất và tiến hành bảo trì định kỳ để duy trì hiệu suất ổn định.
- Thu thập phản hồi và cải tiến liên tục:** Thu thập phản hồi từ người dùng để liên tục cải thiện chất lượng đề xuất và trải nghiệm người dùng. Sử dụng phản hồi này để điều chỉnh và cải thiện mô hình đề xuất.

6. Đánh giá liên tục và cải tiến: Thực hiện đánh giá định kỳ về hiệu suất và độ chính xác của hệ thống. Sử dụng kết quả đánh giá để liên tục cải thiện hệ thống, đảm bảo rằng nó vẫn phù hợp với mục tiêu ban đầu và đáp ứng nhu cầu của người dùng.

Quy trình cơ bản trên đảm bảo sự kết hợp hiệu quả giữa các công nghệ và dịch vụ của Azure để tạo ra một hệ thống đề xuất phim đáng tin cậy và linh hoạt. Sự phối hợp chặt chẽ giữa các thành phần như Azure Databricks, Azure Machine Learning, Azure Storage, SQL Server, SharePoint và Power Apps không chỉ tối ưu hóa quy trình làm việc mà còn cung cấp khả năng mở rộng và thích nghi với các yêu cầu thay đổi.

CHƯƠNG 2. GIỚI THIỆU VỀ HỆ THỐNG AZURE

2.1. Giới thiệu Microsoft Azure



Hình 2.1. Microsoft Azure

Microsoft Azure là giải pháp điện toán đám mây toàn diện được sử dụng để xây dựng, triển khai và quản lý các ứng dụng thông qua mạng lưới trung tâm dữ liệu toàn cầu của Microsoft; giúp ta xây dựng, quản lý và triển khai hiệu quả từ các ứng dụng di động đơn giản đến các giải pháp có quy mô lớn.

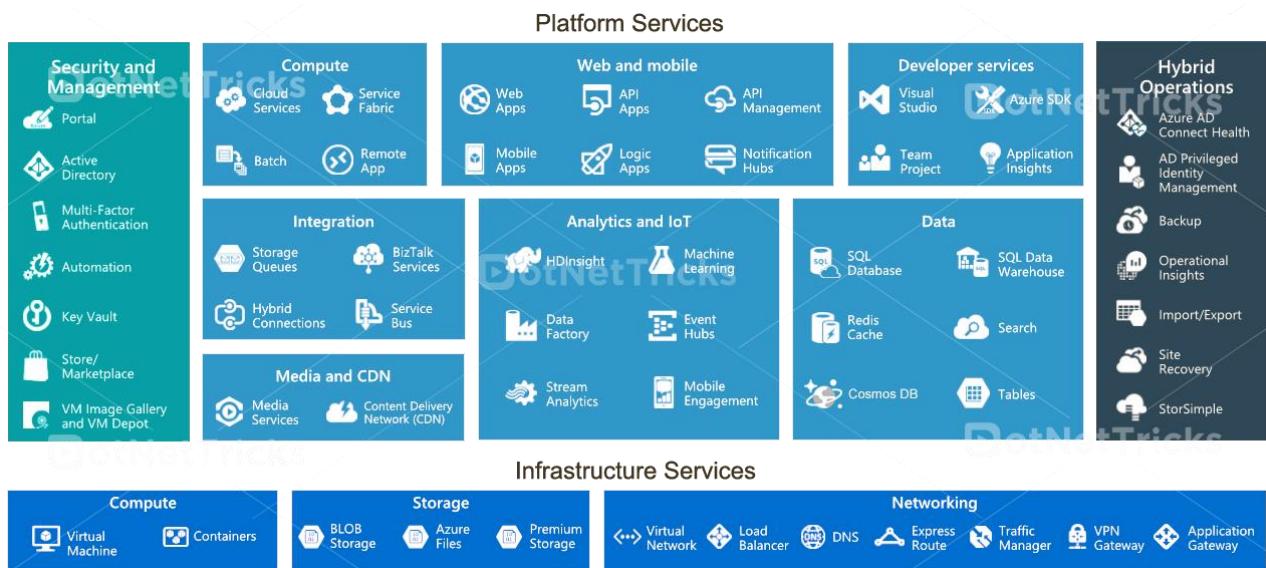
Một số đặc điểm của Azure như:

- Cải thiện công cụ và tự động hóa: Các công cụ tùy chỉnh và quản lý tài nguyên Azure (ARM), giúp đơn giản hóa hoạt động và luôn đảm bảo tuân thủ bằng phương pháp hiện đại nhất.
- Kết nối đa môi trường làm việc: Microsoft Azure cho phép kết nối mạng của doanh nghiệp với đám mây Microsoft hoặc bất kỳ dịch vụ nào khác mà bạn muốn thông qua kết nối riêng tư.
- Azure có nền tảng bảo mật tốt: Bảo mật được cung cấp trong ngành công nghiệp đám mây, nhưng cách tiếp cận chủ động của Azure về bảo mật, tuân thủ và quyền riêng tư là duy nhất. Microsoft dẫn đầu ngành trong việc thiết lập và luôn đáp ứng các yêu cầu về bảo mật và quyền riêng tư rõ ràng.
- Azure hiện có mặt trên toàn cầu: Với các trung tâm dữ liệu ở nhiều khu vực hơn bất kỳ nhà cung cấp đám mây nào khác, Azure cung cấp phạm vi toàn cầu với sự hiện diện địa phương mà nhiều doanh nghiệp và tổ chức cần, cho phép họ giảm chi phí,

thời gian và sự phức tạp của việc vận hành cơ sở hạ tầng toàn cầu trong khi đáp ứng nhu cầu cung cấp dữ liệu cụ thể.

- Những lợi thế của Azure so với AWS: Các tổ chức trên toàn thế giới công nhận Microsoft Azure đã vượt qua Amazon Web Services (AWS) để trở thành dịch vụ điện toán đám mây đáng tin cậy nhất cho doanh nghiệp và cơ sở hạ tầng dạng hybrid vì nhiều lý do: Giá cả cạnh tranh, Được đánh giá cao với mã nguồn mở trên Azure, Tăng cường bảo mật và tuân thủ chủ động, Nhận thêm giá trị từ khoản đầu tư vào Microsoft.

Azure cung cấp hơn 200 dịch vụ (services), được chia thành 18 loại (categories) bao gồm Computing, Networking, Storage, IoT, Migration, Mobile, Analytics, Containers, Artificial Intelligence, Machine Learning, Integration, Management Tools, Developer Tools, Security, Databases, DevOps, Media Identity và Web Services.



Hình 2.2. Các dịch vụ Azure cung cấp

Ngoài ra, Azure cung cấp bốn hình thức điện toán đám mây khác nhau: cơ sở hạ tầng dưới dạng dịch vụ (IaaS), nền tảng dưới dạng dịch vụ (PaaS), phần mềm dưới dạng dịch vụ (SaaS) và các chức năng phi máy chủ.

Microsoft tính phí Azure trên cơ sở thanh toán theo mức sử dụng (PAYG), nghĩa là người đăng ký nhận được hóa đơn mỗi tháng chỉ tính phí cho các tài nguyên và dịch vụ cụ thể mà họ đã sử dụng.

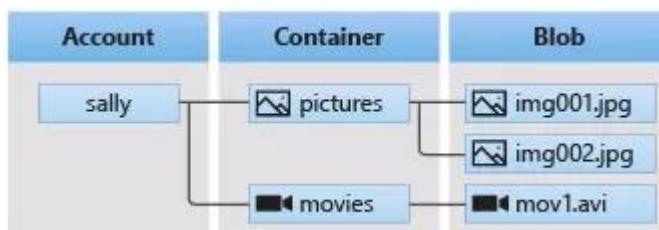
2.2. Azure Blob Storage



Hình 2.3. Microsoft Azure Blob Storage

Azure Blob Storage là một dịch vụ hay đơn giản là một công cụ cho phép lưu trữ dữ liệu phi cấu trúc trên cloud. Mỗi dữ liệu đưa lên để lưu trữ thì ta coi đó như một object, có thể là text, dữ liệu nhị phân, các document hay media file, hoặc là các file cài đặt ... Blob storage hay còn được gọi là Object storage. Azure Blob Storage là NON-SQL Database.

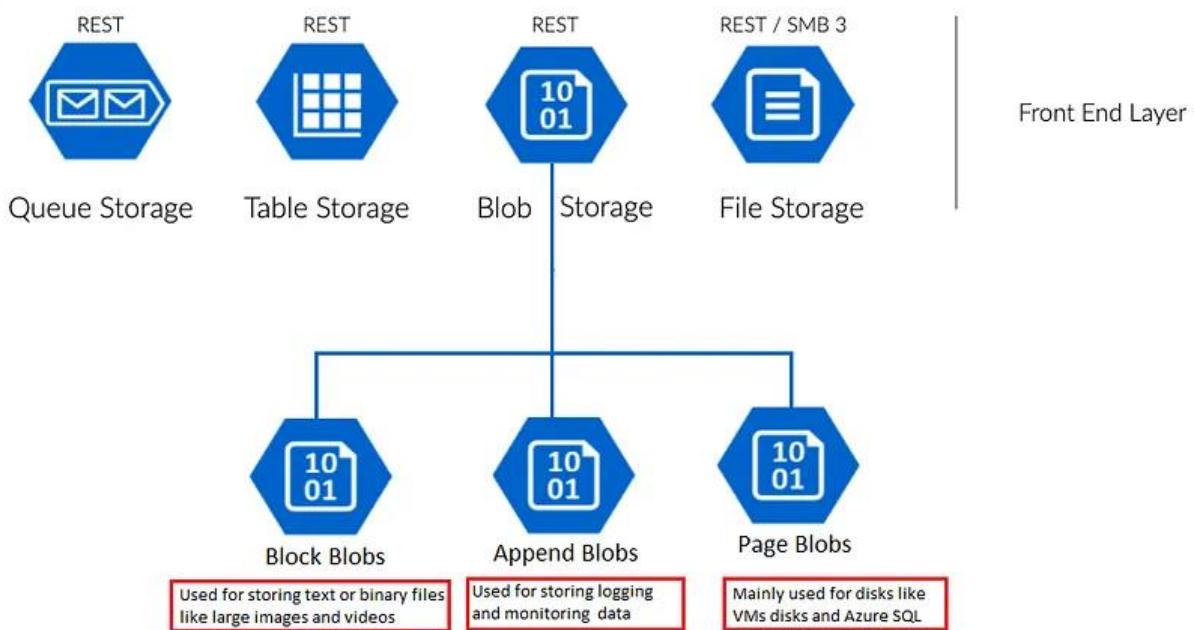
Azure Blob Storage lưu trữ các file tương tự như ta lưu trữ dạng thư mục - file. Mỗi object lưu trữ trên Azure Blob Storage còn được gọi là 1 **blob**. Mỗi **blob** sẽ nằm trong 1 **container** giống như 1 file nằm trong 1 thư mục (được gọi là Storage container). Và các thư mục này có thể nằm trong các thư mục cha khác. Và tất cả chúng thì nằm trong **Storage Account**.



Hình 2.4. Minh họa mối quan hệ giữa các resources

- Storage Account: vì Azure Storage có rất nhiều người dùng nên để lấy được các blob của chúng ta thì ta thông qua Storage Account để xác định phạm vi mà resources của chúng ta được lưu trữ.
- Container: container nằm dưới storage account và chứa các blobs. Tất cả blob phải ở trong 1 container nào đó. 1 Storage account có thể chứa không giới hạn các container. Container có thể chứa không giới hạn các blob. Khi tạo container thì lưu ý tên container phải là tất cả chữ thường.
- Blob: Azure storage hỗ trợ 3 loại blob
 - + Block blobs: lưu trữ dữ liệu dạng text và dữ liệu nhị phân.
 - + Append blobs: lý tưởng cho việc ghi dữ liệu từ máy ảo
 - + Page blobs: lưu trữ các tệp truy cập ngẫu nhiên có kích thước lên đến 8 TB

Azure Storage Architecture



Hình 2.5. Các loại Blob

Tiêu chí	Block Blobs	Append Blobs	Page Blobs
Loại dữ liệu	Text và dữ liệu nhị phân	Text và dữ liệu nhị phân	Các tệp truy cập ngẫu nhiên
Mục đích sử dụng	Lưu trữ dữ liệu chung, streaming video, backup	Ghi nhật ký, ghi dữ liệu liên tục từ máy ảo	Lưu trữ ổ đĩa ảo cho máy ảo Azure, dữ liệu lớn
Kích thước tối đa	190.7 TB mỗi blob	195 GB mỗi blob	8 TB mỗi blob
Ví dụ	Lưu trữ hình ảnh và video cho trang web, lưu trữ tệp tin lớn	Lưu trữ nhật ký ứng dụng, ghi nhật ký hoạt động của máy ảo	Lưu trữ ổ đĩa ảo cho máy ảo Azure, cơ sở dữ liệu lớn

Bảng 2.1. Bảng so sánh các loại Blobs

2.3. Azure Data Factory



Hình 2.6. Azure Data Factory

Azure Data Factory là một dịch vụ được thiết kế bởi Microsoft cho phép các nhà phát triển tích hợp các nguồn dữ liệu khác nhau. Đây là một nền tảng tương tự như SSIS (SQL Server Integration Services) cho phép ta quản lý dữ liệu cả trên nền tảng on-premises và đám mây. Trong đó, SSIS - SQL Server Integration Services - là một thành phần của phần mềm cơ sở dữ liệu Microsoft SQL cho phép bạn thực hiện quá trình di chuyển dữ liệu.

Azure Data Factory là một dịch vụ được thiết kế bởi Microsoft cho phép các nhà phát triển tích hợp các nguồn dữ liệu khác nhau. Đây là một nền tảng tương tự như SSIS (SQL

Server Integration Services) cho phép ta quản lý dữ liệu cả trên nền tảng on-premises và đám mây. ETL là một loại quá trình tích hợp dữ liệu để cập đến ba giai đoạn riêng biệt nhưng liên kết với nhau. Nó được sử dụng để tổng hợp dữ liệu từ nhiều nguồn khác nhau một cách lặp đi lặp lại để xây dựng data warehouse, data hub hoặc data lake.

Azure Data Factory đã trở thành một công cụ quan trọng trong lĩnh vực cloud computing. Trong hầu hết các dự án, ta sẽ cần thực hiện các hoạt động di chuyển dữ liệu qua các mạng khác nhau (trên nền tảng on-premises và đám mây) và các dịch vụ khác nhau (từ và đến các lưu trữ Azure khác nhau).

2.4. Azure Databricks



Hình 2.7. Azure Databricks

Azure Databricks là một nền tảng phân tích đám mây được xây dựng trên nền Apache Spark. Nó cung cấp một không gian làm việc tương tác cho phép người dùng dễ dàng tạo, quản lý và triển khai các nhiệm vụ xử lý dữ liệu lớn và học máy.

Azure Databricks đơn giản hóa quá trình kỹ thuật dữ liệu, khám phá dữ liệu và đào tạo mô hình bằng cách cung cấp một môi trường cộng tác và tương tác. Nó cung cấp một nền tảng có khả năng mở rộng và đáng tin cậy được thiết kế để xử lý tập dữ liệu lớn và các quy trình làm việc phức tạp.

Một trong những thách thức lớn khi làm việc với tập dữ liệu lớn là quản lý sự phức tạp của pipeline data. Với Azure Databricks, người dùng có thể xây dựng và quản lý các pipeline phức tạp bằng cách sử dụng nhiều ngôn ngữ lập trình khác nhau, bao gồm Python, Scala và R.

Databricks cung cấp một giao diện thống nhất giúp dễ dàng quản lý các nhiệm vụ nạp dữ liệu, biến đổi dữ liệu và phân tích và theo dõi hiệu suất của pipeline data.

Không gian làm việc Azure Databricks cung cấp một giao diện và công cụ thống nhất cho hầu hết các nhiệm vụ dữ liệu, bao gồm:

- Lên lịch và quản lý quy trình xử lý dữ liệu
- Tạo bảng điều khiển và trực quan hóa
- Quản lý bảo mật, quản trị, sẵn sàng cao và khả năng phục hồi sau thảm họa
- Khám phá dữ liệu, chú thích và khám phá
- Mô hình hóa, theo dõi và phục vụ mô hình học máy (ML)
- Giải pháp trí tuệ nhân tạo sinh ra.

2.5. Azure Machine Learning Service



Hình 2.8. Azure Machine Learning Service

Microsoft Azure Machine Learning là một dịch vụ dựa trên đám mây cho phép các nhà phát triển và nhà khoa học dữ liệu xây dựng, đào tạo và triển khai các mô hình machine learning một cách dễ dàng.

Nó cung cấp một bộ công cụ và dịch vụ toàn diện phục vụ cho toàn bộ vòng đời máy học, từ chuẩn bị dữ liệu và kỹ thuật tính năng đến đào tạo, đánh giá và triển khai mô hình.

Bằng cách cung cấp một nền tảng tập trung để quản lý và giám sát các dự án máy học, Azure Machine Learning giúp giảm đáng kể sự phức tạp và thời gian liên quan đến việc phát triển và triển khai các giải pháp máy học.

Một trong những tính năng chính của Azure Machine Learning là hỗ trợ máy học tự động hoặc AutoML. AutoML là một công nghệ tiên tiến tự động hóa quy trình chọn thuật toán máy học tốt nhất và điều chỉnh các siêu tham số của nó cho một tập dữ liệu và vấn đề nhất định. Điều này không chỉ tiết kiệm thời gian và công sức mà còn giúp đạt được hiệu suất mô hình tốt hơn bằng cách khám phá phạm vi thuật toán và cấu hình rộng hơn so với khả năng thực hiện thông qua thử nghiệm thủ công.

2.6. Azure Cosmos DB



Hình 2.9. Azure Cosmos DB

Azure Cosmos DB là cơ sở dữ liệu NoSQL và quan hệ được quản lý hoàn toàn dành cho phát triển ứng dụng hiện đại, bao gồm AI, thương mại điện tử, Internet vạn vật, quản lý đặt phòng và các loại giải pháp khác. Azure Cosmos DB cung cấp thời gian phản hồi mili giây một chữ số, khả năng mở rộng tự động và tức thời, cùng với tốc độ được đảm bảo ở mọi quy mô. Tính liên tục của doanh nghiệp được đảm bảo với tính khả dụng được hỗ trợ bởi SLA và bảo mật cấp doanh nghiệp.

Phát triển ứng dụng nhanh hơn và hiệu quả hơn nhờ:

- Phân phối dữ liệu đa khu vực trên toàn cầu một cách sẵn sàng.

- Các API mã nguồn mở.
- Các Bộ phát triển phần mềm (SDKs) cho các ngôn ngữ phổ biến.
- Các chức năng cơ sở dữ liệu AI như tìm kiếm vector bản địa hoặc tích hợp liền mạch với Dịch vụ AI của Azure để hỗ trợ Tạo sinh dữ liệu tăng cường truy xuất.

Azure Cosmos DB là dịch vụ được quản lý hoàn toàn, Azure Cosmos DB giúp bạn thoát khỏi việc quản trị cơ sở dữ liệu với quản lý tự động, cập nhật và vá lỗi. Nó cũng xử lý quản lý dung lượng với các tùy chọn tự động mở rộng không máy chủ và hiệu quả về chi phí, đáp ứng nhu cầu của ứng dụng để phù hợp với dung lượng với nhu cầu.

Bạn có thể Dùng thử Azure Cosmos DB miễn phí mà không cần đăng ký Azure, miễn phí và không ràng buộc hoặc sử dụng gói miễn phí Azure Cosmos DB để có được tài khoản với 1000 RU/s đầu tiên và 25 GB dung lượng lưu trữ miễn phí.

2.7. Azure Container Instances



Hình 2.10. Azure Container Instances

Các container đang trở thành phương thức ưa thích để đóng gói, triển khai và quản lý các ứng dụng đám mây. Azure Container Instances cung cấp cách nhanh nhất và đơn giản nhất để chạy container trong Azure, mà không cần quản lý bất kỳ máy ảo nào và không cần áp dụng dịch vụ cấp cao hơn.

Azure Container Instances là một giải pháp tuyệt vời cho bất kỳ kịch bản nào có thể hoạt động trong các container cô lập, bao gồm các ứng dụng đơn giản, tự động hóa tác vụ và các công việc xây dựng. Đối với các kịch bản mà bạn cần quản lý container đầy đủ, bao

gồm khám phá dịch vụ trên nhiều container, tự động mở rộng và nâng cấp ứng dụng được phối hợp, chúng tôi khuyên bạn nên sử dụng Azure Kubernetes Service (AKS).

Các container cung cấp lợi ích khởi động đáng kể so với các máy ảo (VM). Azure Container Instances có thể khởi động container trong Azure trong vài giây, mà không cần phải cung cấp và quản lý VM.

Azure Container Instances cho phép bạn hiển thị trực tiếp các nhóm container của mình lên internet với địa chỉ IP và tên miền đầy đủ điều kiện (FQDN) và cũng hỗ trợ thực thi một lệnh trong container đang chạy bằng cách cung cấp shell tương tác để hỗ trợ phát triển ứng dụng và khắc phục sự cố. Quyền truy cập diễn ra qua HTTPS, sử dụng TLS để bảo mật kết nối của khách hàng.

Azure Container Instances (ACI) cung cấp môi trường an toàn và tuân thủ quy định để triển khai ứng dụng container. Các tính năng chính bao gồm:

- Bảo mật: ACI cung cấp cùng mức độ bảo mật như VM, bao gồm cô lập ứng dụng, bảo vệ dữ liệu và kiểm soát truy cập.
- Tính linh hoạt: ACI cho phép bạn triển khai các container Linux và Windows, chạy các khối lượng công việc có thể ngắt kết nối và sử dụng GPU NVIDIA.
- Tính kinh tế: ACI cho phép bạn chỉ trả tiền cho các tài nguyên bạn cần và cung cấp các mức giá giảm giá cho các khối lượng công việc có thể ngắt kết nối.

2.8. Azure Kubernetes Service



Azure Kubernetes Service (AKS)

Hình 2.11. Azure Kubernetes Service

Azure Kubernetes Service (AKS) đơn giản hóa việc triển khai cụm Kubernetes được quản lý trong Azure bằng cách chuyển tải chi phí vận hành cho Azure. Là dịch vụ Kubernetes được lưu trữ, Azure xử lý các tác vụ quan trọng, chẳng hạn như giám sát trạng thái và bảo trì. Khi bạn tạo cụm AKS, một mặt phẳng điều khiển sẽ tự động được tạo và cấu hình. Mặt phẳng điều khiển này được cung cấp miễn phí như một tài nguyên Azure được quản lý, được trừu tượng hóa khỏi người dùng. Bạn chỉ phải trả tiền và quản lý các nút được gắn vào cụm AKS.

Khi bạn triển khai cụm AKS, bạn sẽ chỉ định số lượng và kích thước của các nút và AKS sẽ triển khai và cấu hình mặt phẳng điều khiển và các nút Kubernetes. Mạng nâng cao, tích hợp Microsoft Entra, giám sát và các tính năng khác có thể được cấu hình trong quá trình triển khai.

Dưới đây là một số chức năng cụ thể của AKS:

- Tạo và triển khai cluster Kubernetes: AKS cung cấp một cách đơn giản để tạo và triển khai cluster Kubernetes. Doanh nghiệp có thể chỉ cần chọn loại máy ảo, số lượng máy ảo và khu vực Azure.
- Quản lý cluster Kubernetes: AKS cung cấp một bảng điều khiển trực quan để quản lý cluster Kubernetes. Doanh nghiệp có thể sử dụng bảng điều khiển này để thực hiện các tác vụ như triển khai ứng dụng, quản lý dịch vụ và cấu hình cluster.
- Cân bằng tải: AKS tích hợp với Azure Load Balancer để cân bằng tải cho các ứng dụng container. Doanh nghiệp có thể sử dụng Azure Load Balancer để phân phối lưu lượng truy cập đến các ứng dụng container của mình một cách hiệu quả.
- Cập nhật ứng dụng: AKS cho phép doanh nghiệp cập nhật ứng dụng container một cách an toàn và không gián đoạn. AKS sẽ tự động triển khai các bản cập nhật ứng dụng mới lên cluster Kubernetes.
- Mở rộng quy mô: AKS cho phép doanh nghiệp mở rộng quy mô các ứng dụng container theo nhu cầu. Doanh nghiệp có thể sử dụng Azure Autoscale để tự động mở rộng quy mô cluster Kubernetes khi cần thiết.

- Bảo mật: AKS cung cấp các tính năng bảo mật tích hợp để giúp doanh nghiệp bảo vệ các ứng dụng container của mình. Các tính năng này bao gồm xác thực, ủy quyền, mã hóa và giám sát bảo mật.

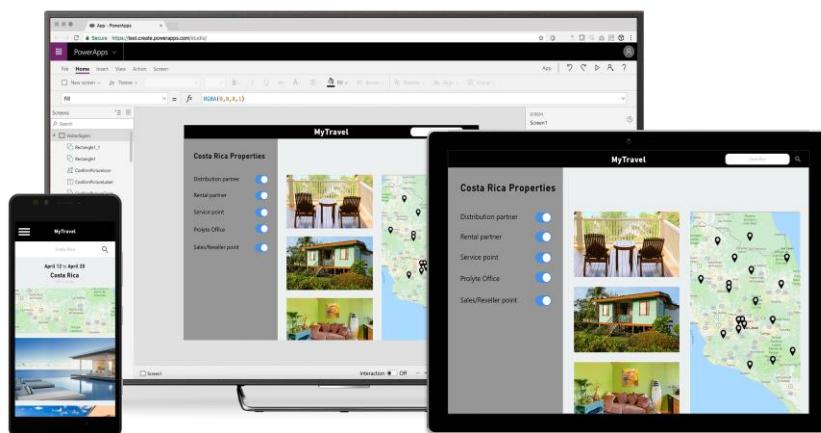
AKS là một dịch vụ mạnh mẽ và linh hoạt giúp doanh nghiệp dễ dàng triển khai, quản lý và bảo mật các ứng dụng container trên Azure.

2.9. PowerApps



Hình 2.12. Power Apps

Power Apps là một bộ sưu tập các ứng dụng, dịch vụ và trình kết nối, cũng như một nền tảng dữ liệu, cung cấp một môi trường phát triển nhanh chóng để xây dựng các ứng dụng tùy chỉnh cho nhu cầu kinh doanh của bạn. Sử dụng Power Apps, bạn có thể nhanh chóng xây dựng các ứng dụng kinh doanh tùy chỉnh kết nối với dữ liệu của bạn được lưu trữ trong nền tảng dữ liệu cơ bản (Microsoft Dataverse) hoặc trong các nguồn dữ liệu trực tuyến và cục bộ khác (chẳng hạn như SharePoint, Microsoft 365, Dynamics 365, SQL Server, v.v.).



Hình 2.13. Ứng dụng có thể chạy trên trình duyệt và di động

Ứng dụng được xây dựng bằng Power Apps cung cấp logic kinh doanh phong phú và khả năng quản lý quy trình công việc để chuyển đổi các hoạt động kinh doanh thủ công của bạn thành quy trình kỹ thuật số, tự động hóa. Hơn nữa, các ứng dụng được xây dựng bằng Power Apps có thiết kế đáp ứng và có thể chạy liền mạch trên trình duyệt và trên thiết bị di động (điện thoại hoặc máy tính bảng). Power Apps "dân chủ hóa" trải nghiệm xây dựng ứng dụng kinh doanh bằng cách cho phép người dùng tạo các ứng dụng kinh doanh tùy chỉnh phong phú mà không cần viết mã.

Power Apps cũng cung cấp nền tảng mở rộng cho phép các nhà phát triển chuyên nghiệp tương tác lập trình với dữ liệu và siêu dữ liệu, áp dụng logic kinh doanh, tạo trình kết nối tùy chỉnh và tích hợp với dữ liệu bên ngoài.

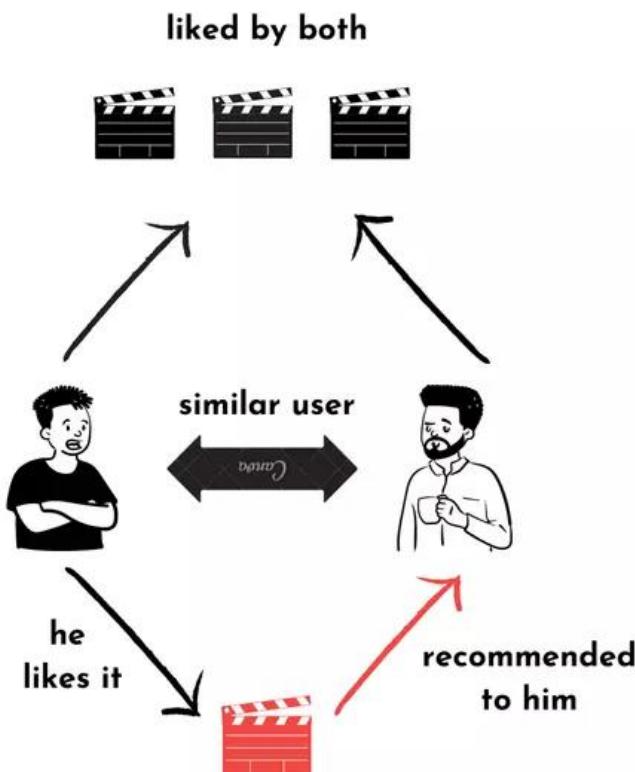
Power Apps có nhiều ưu điểm, bao gồm:

- Dễ sử dụng: Power Apps sử dụng giao diện kéo và thả đơn giản, giúp người dùng không có kinh nghiệm về lập trình cũng có thể tạo các ứng dụng phức tạp.
- Tùy biến cao: Power Apps cho phép người dùng tùy chỉnh ứng dụng theo nhu cầu cụ thể của mình.
- Linh hoạt: Power Apps có thể được sử dụng để tạo các ứng dụng trên nhiều nền tảng, bao gồm web, máy tính để bàn và thiết bị di động.
- Khả năng tích hợp: Power Apps có thể được tích hợp với các ứng dụng và dịch vụ khác của Microsoft, cũng như với các ứng dụng và dịch vụ bên ngoài.
- Dùng thử Power Apps miễn phí: Bạn có thể xây dựng Power Apps miễn phí. Chỉ cần đăng nhập vào Power Apps.

CHƯƠNG 3. QUY TRÌNH RA QUYẾT ĐỊNH

3.1. Phương pháp đề xuất phim.

Thuật toán Collaborative Filtering (CF) hoạt động dựa trên ý tưởng cơ bản là dự đoán mức độ yêu thích của một người dùng (user) đối với một sản phẩm (item) dựa trên sở thích của những người dùng khác có đặc điểm tương tự. Độ “giống nhau” giữa các người dùng được xác định dựa vào mức độ quan tâm (rating) mà họ đưa ra cho các sản phẩm khác trong quá khứ. Ví dụ, nếu hai người dùng A và B đều thích các phim cảnh sát hình sự, và B thích phim “Người phán xử”, thì có khả năng cao A cũng sẽ thích phim này. Dựa vào đó, hệ thống sẽ đề xuất “Người phán xử” cho A.



Hình 3.1. Gợi ý phim dựa trên sở thích của người dùng

Câu hỏi đặt ra cho bài toán Collaborative filtering là:

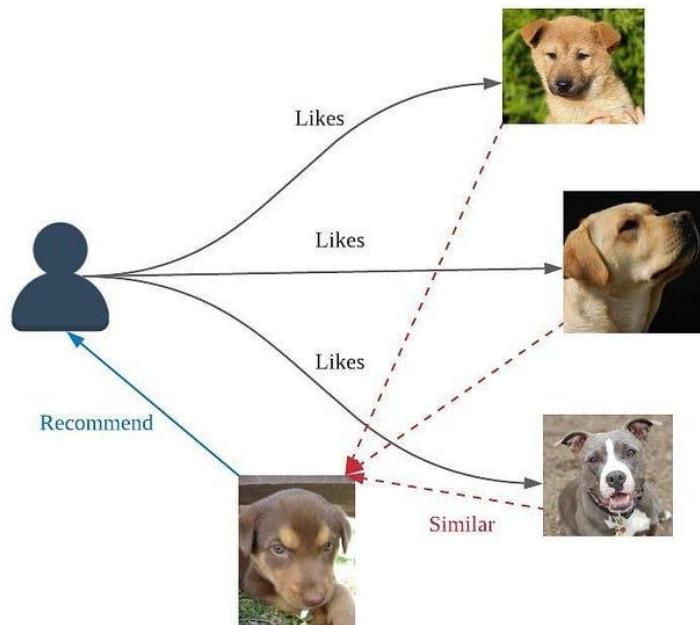
- Làm thế nào xác định được sự giống nhau giữa hai users?

- Khi đã xác định được các users gần giống nhau (similar users) rồi, làm thế nào dự đoán được mức độ quan tâm của một user lên một item?

Về cơ bản Collaborative Filtering được chia làm 2 phương pháp:

- Memory-Based: hay còn gọi là thuật toán Collaborative Filtering dựa trên vùng lân cận. Các vùng lân cận này có thể được xác định theo một trong hai cách
 - + User – User Collaborative Filtering: Ý tưởng cơ bản là phân chia các User tương tự nhau vào chung một nhóm. Nếu một User bất kỳ trong nhóm thích một Item nào đó thì item đó sẽ được đề xuất cho toàn bộ các Users khác trong nhóm đó.
 - + Item-Item Collaborative Filtering: Ý tưởng cơ bản là phân chia Items tương tự nhau vào chung một nhóm. Nếu một User thích một Item nào trong nhóm đó thì tất cả Item còn lại trong cùng nhóm sẽ được đề xuất cho User đó.

Trong bài toán nào nhóm sẽ phân tích kĩ vào hướng Item-Item Collaborative Filtering.



Hình 3.2. Item-based collaborative filtering

Phương pháp tiếp cận của Item-Item Collaborative Filtering được hiện thực như sau:

- Biểu diễn vector cho mỗi item: Mỗi sản phẩm (item) được biểu diễn dưới dạng một vector thuộc tính. Điều này cho phép hệ thống hiểu và so sánh các sản phẩm với nhau.
- Tính toán độ tương đồng giữa các item: Dựa trên các vector thuộc tính, hệ thống tính toán mức độ tương đồng giữa các sản phẩm. Điều này giúp xác định những sản phẩm tương tự nhau.
- Xác định độ yêu thích của người dùng user (U) đối với một item (I): Hệ thống chọn ra k sản phẩm mà người dùng U đã đánh giá và có độ tương đồng cao nhất với sản phẩm I. Dựa trên phản hồi của người dùng U đối với k sản phẩm này, hệ thống tính toán xem người dùng có thể yêu thích sản phẩm I đến mức nào.
- Gợi ý các sản phẩm: Cuối cùng, hệ thống chọn ra những sản phẩm mà dự đoán người dùng U sẽ yêu thích nhất để đưa ra gợi ý.

Ví dụ về utility matrix dựa trên số sao một user rate cho một item. Như Utility Matrix dưới đây thể hiện mức độ yêu thích của User đối với mỗi item, tiến hành chuẩn hóa và điền các giá trị thiếu cho matrix dưới theo Item-item Collaborative Filtering

	u_0	u_1	u_2	u_3	u_4	u_5	u_6
i_0	5	5	2	0	1	?	?
i_1	3	?	?	0	?	?	?
i_2	?	4	1	?	?	1	2
i_3	2	2	3	4	4	?	4
i_4	2	0	4	?	?	?	5

Hình 3.3. Ví dụ về CF

Các bước để tìm bộ phim tương đồng với item-item Collaborative Filtering

- Tính trung bình cộng ratings của các items

- Ma trận Normalized Utility Matrix bằng cách: Thực hiện chuẩn hoá bằng cách trừ các rating đã biết của item cho trung bình cộng rating tương ứng của item đó và thay các ratings chưa biết bằng 0.
- Tính Similarity bằng similarity function cho các item.
- Sử dụng Similarity matrix và normalized utility matrix để dự đoán ra rating của các users với mỗi items.
- Ví dụ: để sự đoán rating của User U cho item I
 - + Bước 1: Tìm tập $N(i, u)$ mà các items U đã đánh giá.
 - + Bước 2: Lấy similarities của I với các items trong tập $N(u, i)$. Chọn ra k items gần nhất (similarity cao nhất) với I .
 - + Bước 3: Tính rating dự đoán theo công thức:

$$\hat{y}_{i,u} = \frac{\sum_{u_j \in N(u,i)} \bar{y}_{i,u_j} \text{sim}(u, u_j)}{\sum_{u_j \in N(u,i)} |\text{sim}(u, u_j)|}$$

Đối với ví dụ trên: chúng ta có thể thực hiện các step tìm ra Normalized utility matrix \bar{Y} như sau:

	u_0	u_1	u_2	u_3	u_4	u_5	u_6
i_0	5	5	2	0	1	?	?
i_1	4	?	?	0	?	2	?
i_2	?	4	1	?	?	1	1
i_3	2	2	3	4	4	?	4
i_4	2	0	4	?	?	?	5

a) Original utility matrix Y and mean item ratings.

	u_0	u_1	u_2	u_3	u_4	u_5	u_6
i_0	2.4	2.4	-6	-2.6	-1.6	0	0
i_1	2	0	0	-2	0	0	0
i_2	0	2.25	-0.75	0	0	-0.75	-0.75
i_3	-1.17	-1.17	-0.17	0.83	0.83	0	0.83
i_4	-0.75	-2.75	1.25	0	0	0	2.25

b) Normalized utility matrix \bar{Y} .

	i_0	i_1	i_2	i_3	i_4
i_0	1	0.77	0.49	-0.89	-0.52
i_1	0.77	1	0	-0.64	-0.14
i_2	0.49	0	1	-0.55	-0.88
i_3	-0.89	-0.64	-0.55	1	0.68
i_4	-0.52	-0.14	-0.88	0.68	1

c) Item similarity matrix S .

	u_0	u_1	u_2	u_3	u_4	u_5	u_6
i_0	2.4	2.4	-6	-2.6	-1.6	-0.29	-1.52
i_1	2	2.4	-0.6	-2	-1.25	0	-2.25
i_2	2.4	2.25	-0.75	-2.6	-1.20	-0.75	-0.75
i_3	-1.17	-1.17	-0.17	0.83	0.83	0.34	0.83
i_4	-0.75	-2.75	1.25	1.03	1.16	0.65	2.25

d) Normalized utility matrix \bar{Y} .

Hình 3.4. Mô phỏng quá trình gợi ý phim theo phương pháp item-item Collaborative Filtering

3.2. Phương pháp đánh giá

Trong mô hình dự đoán các hệ thống gợi ý, người ta thường sử dụng phương pháp căn của sai số bình phương trung bình RMSE (Root mean squared error) và sai số tuyệt đối trung bình MAE (Mean Absolute Error) để đánh giá tính chính xác của các dự đoán.

Sai số tuyệt đối trung bình

$$MAE = \frac{\sum_{i=1}^n |p_{i,j} - r_{i,j}|}{n}$$

Căn của sai số tuyệt đối trung bình

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (p_{i,j} - r_{i,j})^2}{n}}$$

Trong đó: Tính chính xác của các dự đoán được đo trên n quan sát, $p_{i,j}$ là giá trị dự đoán của người dùng i trên bộ phim j ; $r_{i,j}$ là giá trị xếp hạng thực tế.

3.3. Thuật toán Alternating Least Square

Alternating least squares (ALS) hay còn gọi là phương pháp bình phương tối thiểu thay phiên nhau. Gọi là thay phiên nhau bởi hàm mất mát trên có 2 biến ta thay phiên cố định 1 biến rồi tối ưu hàm theo biến còn lại.

Thuật toán ALS là một phương pháp phân tích ma trận được sử dụng trong các hệ thống khuyến nghị. Thuật toán này giúp giảm chiều của ma trận đánh giá bằng cách phân tích ma trận thành hai ma trận con có kích thước nhỏ hơn.

Thuật toán ALS sử dụng phương pháp lặp để tối ưu hàm mất mát. Thuật toán này được sử dụng rộng rãi trong các hệ thống khuyến nghị do tính đơn giản và hiệu quả của nó.

Algorithm 1 ALS for Matrix Completion

Initialize X, Y

repeat

for $u = 1 \dots n$ **do**

$$x_u = \left(\sum_{r_{ui} \in r_{u*}} y_i y_i^\top + \lambda I_k \right)^{-1} \sum_{r_{ui} \in r_{u*}} r_{ui} y_i$$

end for

for $i = 1 \dots m$ **do**

$$y_i = \left(\sum_{r_{ui} \in r_{*i}} x_u x_u^\top + \lambda I_k \right)^{-1} \sum_{r_{ui} \in r_{*i}} r_{ui} x_u$$

end for

until convergence

Hình 3.5. Phương pháp lặp trong ALS

Đây là một thuật toán hỗ trợ trên pyspark và nó chạy song song.

Ví dụ: Ta có 1 ma trận 2 chiều của từng user và bộ phim, trong đó các giá trị 0 nghĩa là người dùng chưa xem bộ phim đó.

		movield			
		111	222	333	444
userId	1	2	2	5	0
	2	0	0	3	1
	3	0	0	0	4
	4	3	0	1	1
	5	0	2	2	0

Hình 3.6. Mô phỏng dữ liệu đánh giá của khách hàng cho từng bộ phim

Công việc của ta dự đoán xem với các bộ phim mà user chưa xem họ có xu hướng rating bộ phim đó bao nhiêu điểm, từ những số điểm đó mà ta có thể dễ dàng gợi ý các bộ phim theo điểm số rating từ cao xuống thấp cho mỗi user.

Nhiệm vụ của model ALS là tìm 2 ma trận, tạm gọi là ma trận U và P sao cho tích vô hướng của chúng sẽ xấp xỉ bằng ma trận ban đầu (ma trận userId và movieId). Khi các ma trận đã được tìm thấy ta có thể dự đoán chỉ số rating của toàn bộ sản phẩm (bộ phim) ứng với mỗi user. Ví dụ bạn muốn dự đoán cho user i và bộ phim j thì ta chỉ cần lấy tích của hàng i trong ma trận U và hàng j trong ma trận P là được.

		movieId			
		111	222	333	444
userId	1	2	2	5	0
	2	0	0	3	1
	3	0	0	0	4
	4	3	0	1	1
	5	0	2	2	0

ALS →

		P matrix			
		111	222	333	444
U matrix	1	2	2	5	5.34007
	2	6.74569	6.23289	3	1
	3	6.48773	7.35803	3.86885	4
	4	3	7.05254	1	1
	5	5.54446	2	2	1.95555

Hình 3.7. Kết quả khi sử dụng ALS

3.4. Thuật toán Surprise Singular Value Decomposition

Thuật toán Surprise SVD là một biến thể nâng cao và tinh chỉnh của phương pháp phân rã ma trận Singular Value Decomposition (SVD) truyền thống, được điều chỉnh đặc biệt cho việc ứng dụng trong hệ thống đề xuất (recommendation systems). Được tích hợp trong thư viện Surprise của Python, thuật toán này nổi bật với khả năng dự đoán đánh giá và sở thích của người dùng một cách chính xác.

Trong SVD truyền thống, trọng tâm là phân rã ma trận đánh giá thành các yếu tố biểu diễn các đặc tính ẩn của người dùng và mặt hàng nhưng Surprise SVD mở rộng điều này bằng cách thêm vào độ lệch người dùng b_u (user biases) và độ lệch mặt hàng b_i (item biases) vào mô hình. Độ lệch người dùng giúp mô tả xu hướng của một người dùng đánh giá cao hoặc thấp hơn mức trung bình, trong khi độ lệch mặt hàng phản ánh mức độ phổ biến hoặc ưa chuộng của mặt hàng đó. Điều này rất quan trọng trong việc nắm bắt sở thích cá nhân và đặc tính cụ thể của mặt hàng.

Mô hình SVD dự đoán đánh giá dựa trên công thức:

$$\hat{r}_{ui} = \mu + b_u + b_i + q_i^T p_u$$

Trong đó: μ là đánh giá trung bình toàn cầu trong tập dữ liệu.

b_u là độ lệch người dùng u .

b_i là độ lệch mặt hàng i .

q_i và p_u q_i là các vectơ nhân tố ẩn cho mặt hàng i và người dùng u .

$q_i^T p_u$ là tích vô hướng của các yếu tố ẩn này, nắm bắt sự tương tác giữa sở thích của người dùng và thuộc tính của mặt hàng.

Mô hình SVD sử dụng một cách tiếp cận tối ưu hóa có điều kiện phạt (regularized optimization problem) để cân bằng giữa việc khớp dữ liệu và ngăn chặn overfitting. Trong tối ưu hóa có điều kiện phạt của mô hình SVD, quá trình này tập trung vào việc giảm thiểu tổng bình phương của sai số giữa đánh giá thực tế r_{ui} và đánh giá dự đoán \hat{r}_{ui} . Cụ thể, công thức cho vấn đề tối ưu hóa trong SVD như sau:

$$\sum (r_{ui} - (\mu + b_u + b_i + q_i^T p_u))^2 + \lambda(\|q_i\|^2 + \|p_u\|^2 + b_i^2 + b_u^2)$$

Trong đó: r_{ui} : Là đánh giá thực tế của người dùng u đối với mặt hàng i .

$\hat{r}_{ui} = \mu + b_u + b_i + q_i^T p_u$: Là đánh giá dự đoán của mô hình SVD có công thức đã giới thiệu ở trên.

λ : Là tham số điều chỉnh (regularization parameter), giúp cân bằng giữa việc khớp dữ liệu và tránh overfitting.

$\|q_i\|^2, \|p_u\|^2, b_i^2, b_u^2$: Là các thuật ngữ điều chỉnh, đại diện cho tổng bình phương độ lớn của các vectơ nhân tố và độ lệch. Điều này đảm bảo rằng mô hình không quá phụ thuộc vào bất kỳ yếu tố cụ thể nào, giảm thiểu nguy cơ overfitting.

Thuật toán SVD sử dụng Stochastic Gradient Descent (SGD) để cập nhật các tham số. SGD là một thuật toán rất phổ biến để tối ưu hóa trong đó các tham số (ở đây là độ lệch và vectơ nhân tố) được tăng dần theo độ dốc âm với hàm tối ưu hóa. Thuật toán về cơ bản thực hiện các bước sau cho một số lần lặp nhất định:

$$b_u \leftarrow b_u + \gamma(e_{ui} - \lambda b_u)$$

$$b_i \leftarrow b_i + \gamma(e_{ui} - \lambda b_i)$$

$$p_u \leftarrow p_u + \gamma(e_{ui} \cdot q_i - \lambda p_u)$$

$$q_i \leftarrow q_i + \gamma(e_{ui} \cdot p_u - \lambda q_i)$$

Trong đó: γ là tốc độ học, điều khiển tốc độ hội tụ của thuật toán.

$e_{ui} = r_{ui} - \hat{r}_{ui} = r_{ui} - (\mu + b_u + b_i + q_i^T p_u)$ là sai số dự đoán cho cặp (u, i) .

Surprise SVD được sử dụng rộng rãi trong các loại hệ thống đề xuất, từ đề xuất phim, sách, đến sản phẩm. Nó được đánh giá cao về khả năng xử lý dữ liệu thưa thớt và cung cấp dự đoán chính xác.

3.5. Bảng so sánh kết quả

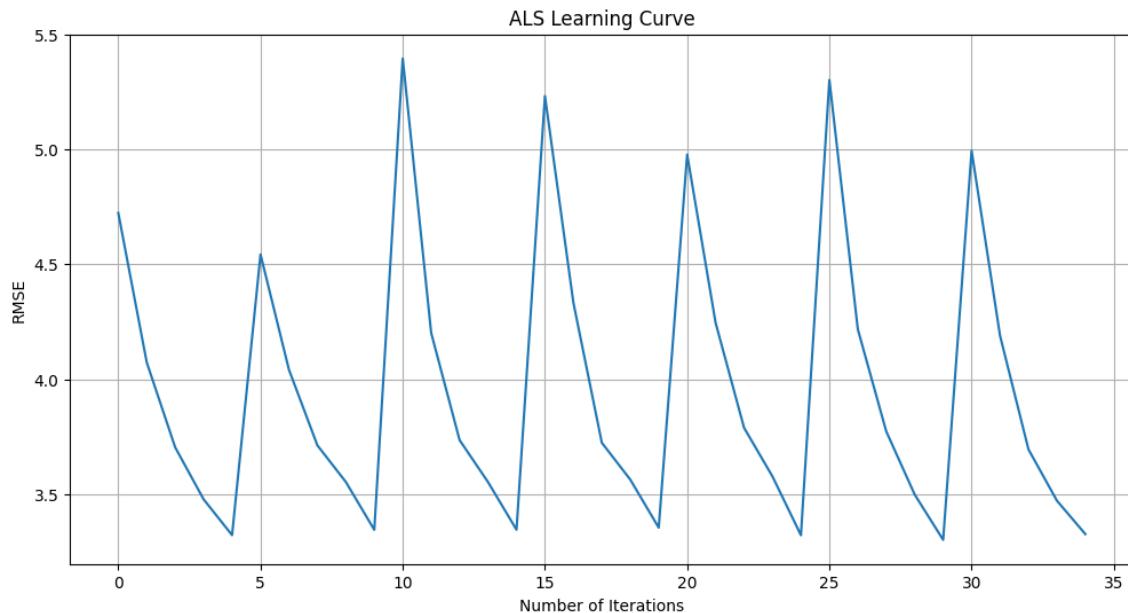
	Thuật toán ALS	Thuật toán Surprise SVD
RMSE	216508.8175	306002.2770

Bảng 3.1. Bảng so sánh kết quả

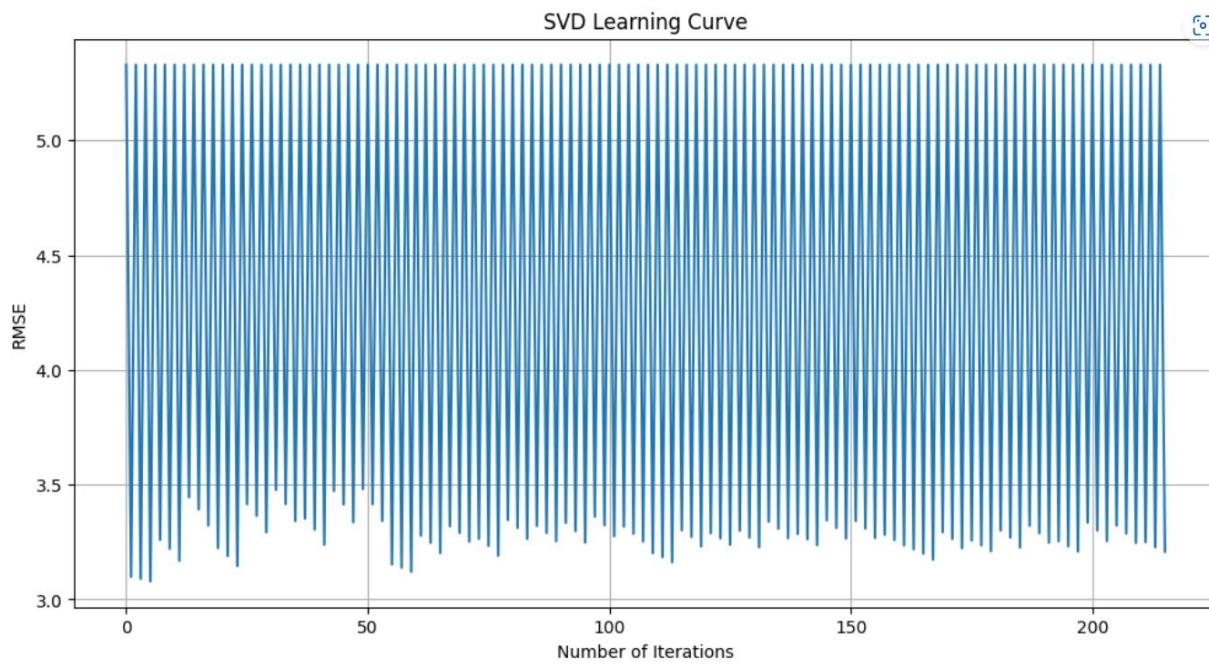
Từ kết quả trên, ta có nhận xét thuật toán ALS có độ chính xác cao hơn SVD. RMSE của ALS nhỏ hơn SVD, điều này cho thấy ALS dự đoán các giá trị đánh giá tốt hơn SVD và ALS phù hợp hơn với dữ liệu của bài toán hơn SVD. Với kết quả trên, nhóm quyết định sử dụng mô hình ALS cho bài toán này để đưa ra các gợi ý phim trong hệ thống gợi ý phim mà nhóm xây dựng.

3.6. Trực quan hóa dữ liệu

Trong phần trực quan hóa, nhóm sẽ tiến hành trực quan các giá trị RMSE trong mỗi lần lặp của quá trình tìm tham số tốt nhất.



Hình 3.8. Thuật toán ALS

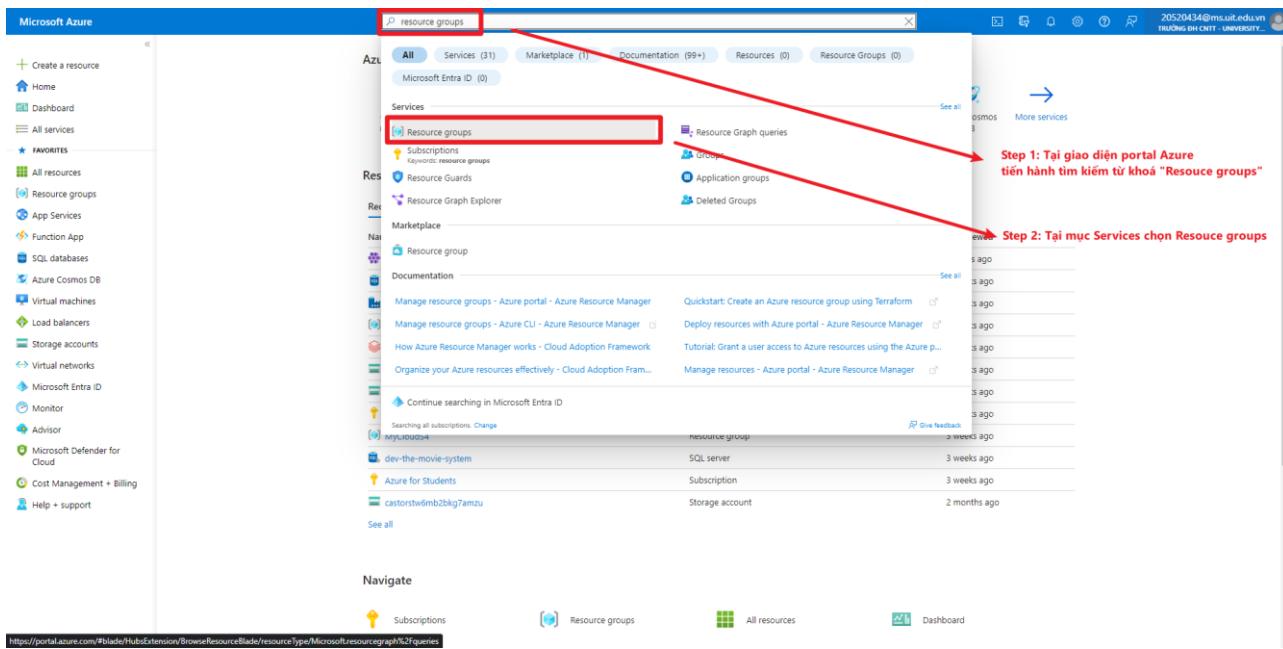


Hình 3.9. Thuật toán SVD

CHƯƠNG 4. CÁC BƯỚC TRIỂN KHAI

4.1. Triển khai môi trường Azure

1. Truy cập và đăng nhập vào [Azure Portal](#) bằng tài khoản Azure của bạn.
2. Tạo Resource Group:
 - Chọn “Resource Group” từ menu hoặc tìm kiếm nó.
 - Nhấn vào “Create” để khởi tạo một Resource Group mới
 - Điền các thông tin:
 - + Resource group name: Đặt tên cho Resource Group của bạn.
 - + Region: Chọn khu vực mà bạn muốn lưu trữ tài nguyên. Hãy chọn khu vực gần nhất với vị trí của bạn hoặc người dùng cuối để tối ưu hóa hiệu suất.



Hình 4.1. Tìm kiếm Service Resource Groups

Step 3: Chọn create để khởi tạo Resource Groups

The screenshot shows the Microsoft Azure Resource Groups page. On the left, there's a sidebar with various service icons. In the center, under 'Resource groups', there's a list of existing resource groups: 'cloud-shell-storage-southeastasia', 'MyCloud54', and 'The-Movie-System'. A red arrow points to the '+ Create' button at the top of the list.

Hình 4.2. Tạo Resource Groups

Tại Basics tiến hành điền các thông tin cần thiết

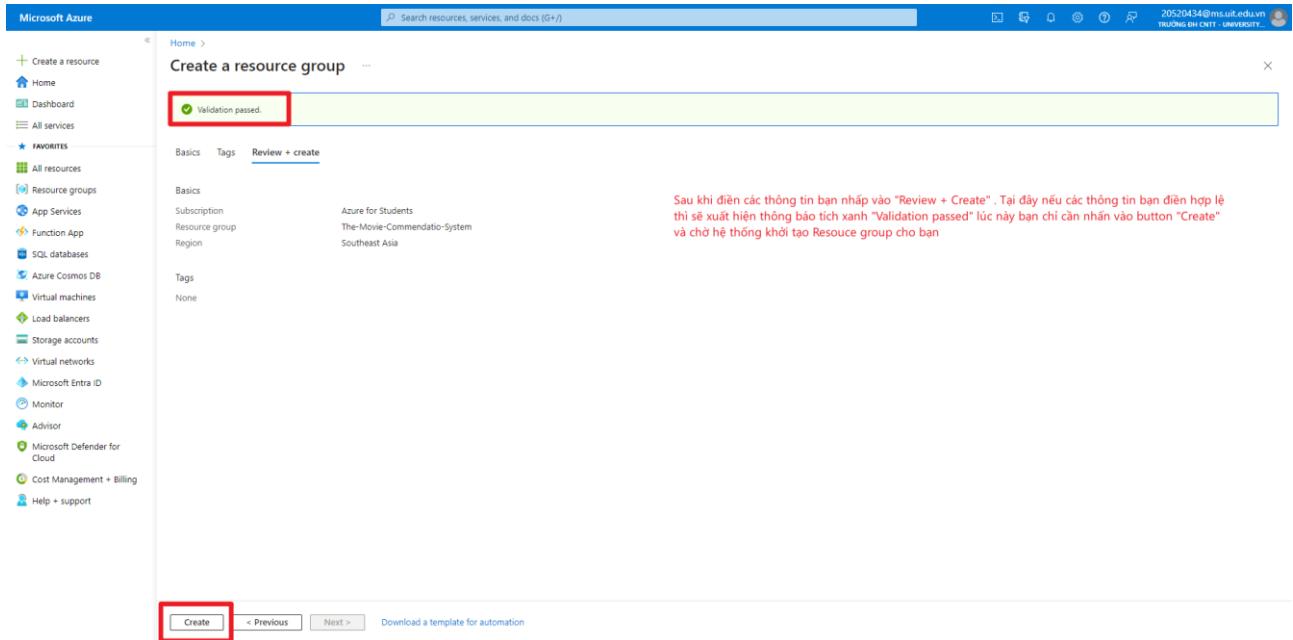
Step 4: Chọn Subscriptions: Chọn subsciptions phù hợp giúp bạn quản lý thông tin chi phí các tài nguyên phát sinh

Step 5: Đặt tên Resouce group: Tiến hành đặt tên resouce group theo quy tắc đặt tên của Azure.

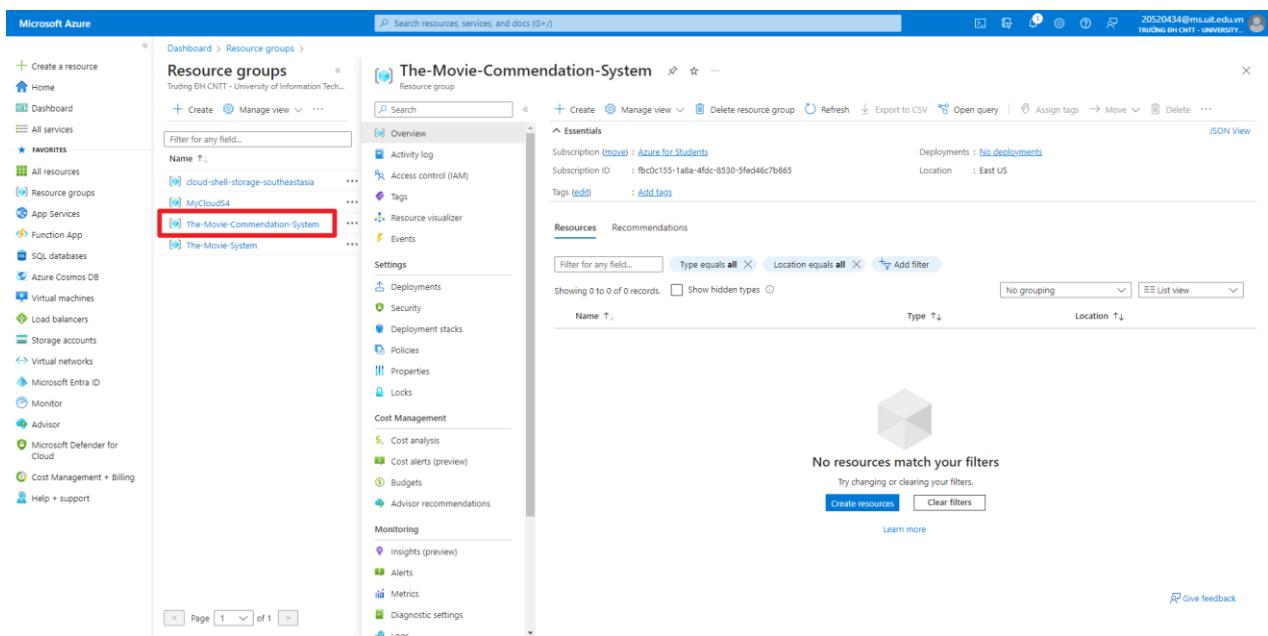
Step 6: Lựa chọn Region: Chọn Region phù hợp và gán bạn giúp trải nghiệm hệ thống và việc lưu trữ mang lại hiệu suất tốt hơn

The screenshot shows the 'Create a resource group' wizard. It has three tabs: Basics, Tags, and Review + create. The Basics tab is active. It includes fields for 'Project details' (Subscription: Azure for Students, Resource group: 'The-Movie-Commandment-System') and 'Resource details' (Region: '(Asia Pacific) Southeast Asia'). Red arrows point from the text descriptions to the corresponding dropdown menus in the UI.

Hình 4.3. Điền các thông tin cho Resource Groups



Hình 4.4. Kiểm tra thông tin trước khi khởi tạo



Hình 4.5. Kết quả khởi tạo Resouce Groups

4.2. Triển khai môi trường Azure Storage

Tiến hành tìm kiếm Services:
Storage account

The screenshot shows the Microsoft Azure search interface. The search bar at the top contains the text "Storage accounts". Below the search bar, there is a list of services under the heading "Services". One item, "Storage accounts", is highlighted with a red box and has a red arrow pointing from the search term above it to this item. Other items in the list include "Storage browser", "Storage movers", "Storage accounts (classic)", "Data Lake Storage Gen1", "Azure Native Qumulo Scalable File Service", "Publisher Artifact Stores", "Storage account", "Backup and Site Recovery", "Azure File Sync", "Data Lake Storage Gen1", "Azure Blob Storage on IoT Edge", "Azure Stack Edge", "NetApp BlueXP", "SFTP Gateway for Azure - SFTP to Blob storage", "Concepts - Storage in Azure Kubernetes Services (AKS) - Azure Ku...", "Tutorial: Connect to a storage account using an Azure Private End...", "Get storage account configuration information - Azure Storage", "Monitor Azure Storage services with Azure Monitor Storage insights", and "Quickstart: Integrate an Azure Storage account with Azure CDN", "Quickstart: Upload, download, and list blobs - Azure CLI - Azure St...", "Use private endpoints - Azure Storage", and "Quickstart: Upload, download, and list blobs - Azure PowerShell - ...". The URL in the address bar is https://portal.azure.com/#blade/Microsoft_Azure_Storage/StorageBrowserAccountPicker.

Hình 4.6. Tìm kiếm Storage Accounts

Nhấn Create

The screenshot shows the Microsoft Azure Storage accounts page. At the top, there is a search bar and a "Create" button highlighted with a red box and a red arrow pointing to it. Below the search bar, there is a table displaying four storage accounts. The columns are: Name, Type, Kind, Resource group, Location, and Subscription. The data in the table is as follows:

Name	Type	Kind	Resource group	Location	Subscription
castorsted6mb2bkg7am	Storage account	Storage	MyCloud54	West US	Azure for Students
masterdatamovie	Storage account	StorageV2	The-Movie-System	East US	Azure for Students
synapseditfrm	Storage account	StorageV2	The-Movie-System	East US	Azure for Students
themoviesystembc7	Storage account	Storage	The-Movie-System	East US	Azure for Students

The URL in the address bar is <https://portal.azure.com/#view/HubExtension/BrowseResource/resourceType/Microsoft.Storage%2fstorageAccounts>.

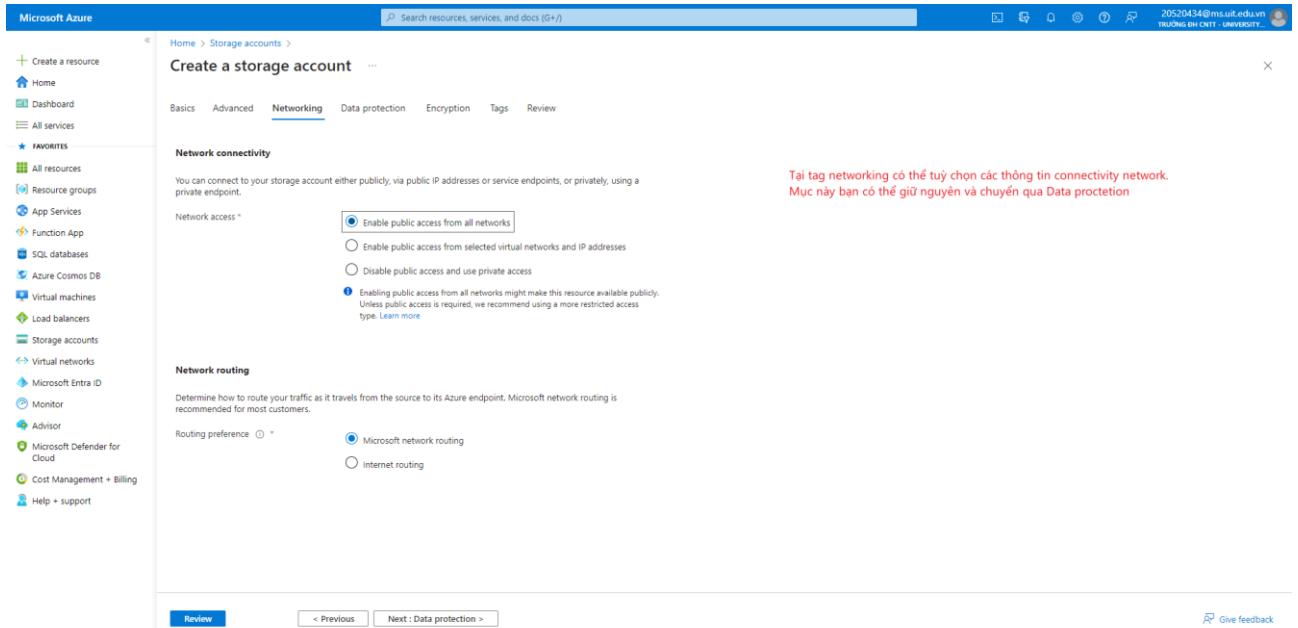
Hình 4.7. Tạo mới Storage Accounts

Tại tag Basics dien các thông tin cơ bản như:
 + Chon thông tin: Subscription và Resouce group
 +Tên storage đưa vào một số quy tắc do Azure quy định cho tên Storage
 + Region phù hợp với nơi bạn sinh sống
 + Performance và Redudancy: Tùy thuộc vào nhu cầu và mục đích sử dụng

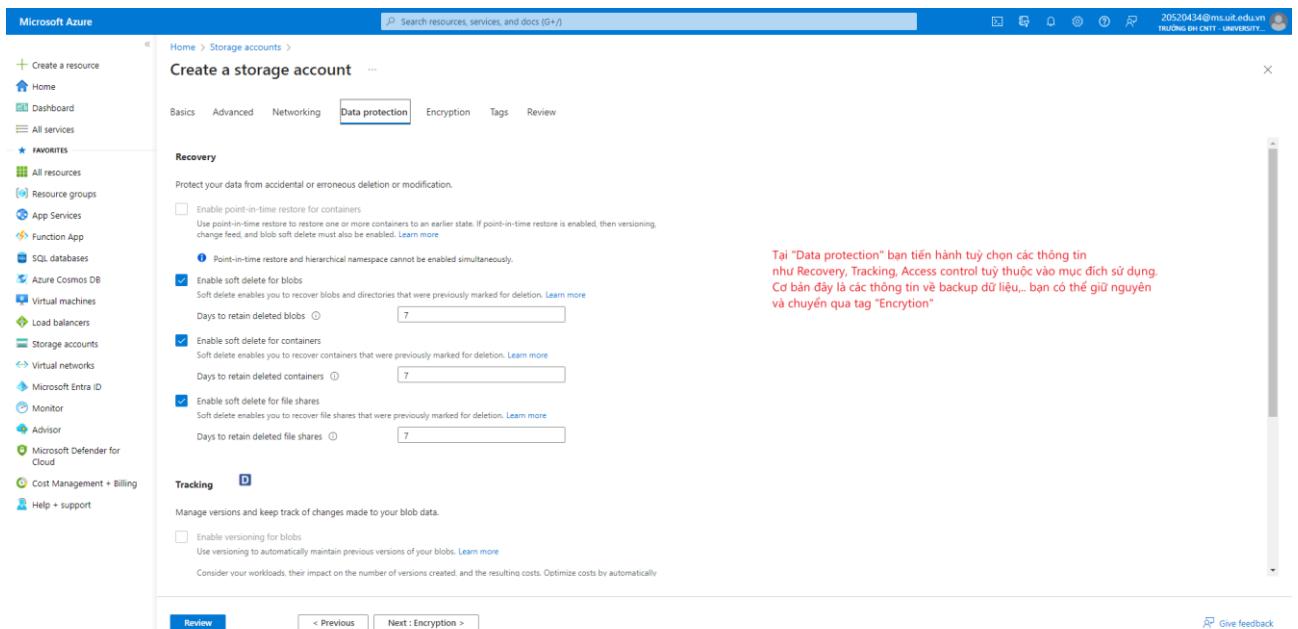
Hình 4.8. Khởi tạo các thông tin cơ bản của Storage tại Base

Tại tag Advanced tiến hành tuỳ chọn các thông tin: Security, Hierarchical Namespace, Access protocols, Blob storage, Azure Files
 tuỳ thuộc vào nhu cầu và mục đích lưu trữ dữ liệu.
 Tại tag này mình chỉ tuỳ chọn thêm Enable hierarchical namespace để tăng tốc khôi lượng công việc phân tích dữ liệu lớn và hỗ trợ danh sách kiểm soát truy cập (ACL)

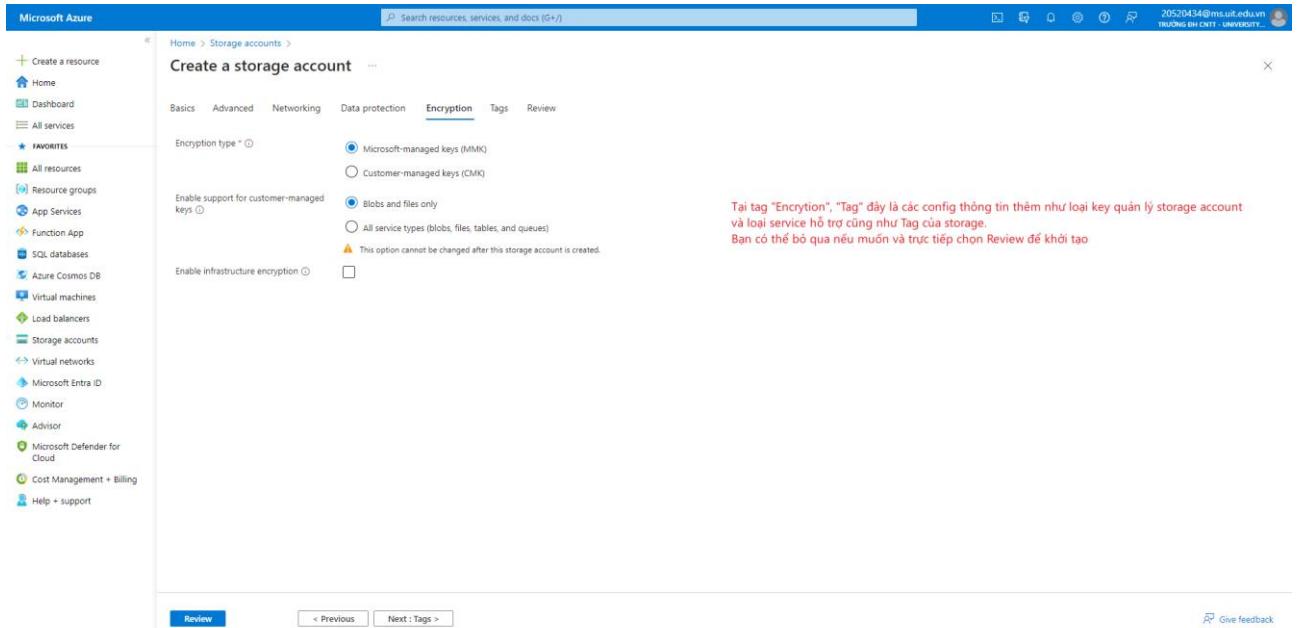
Hình 4.9. Khởi tạo các thông tin cơ bản của Storage tại Advanced



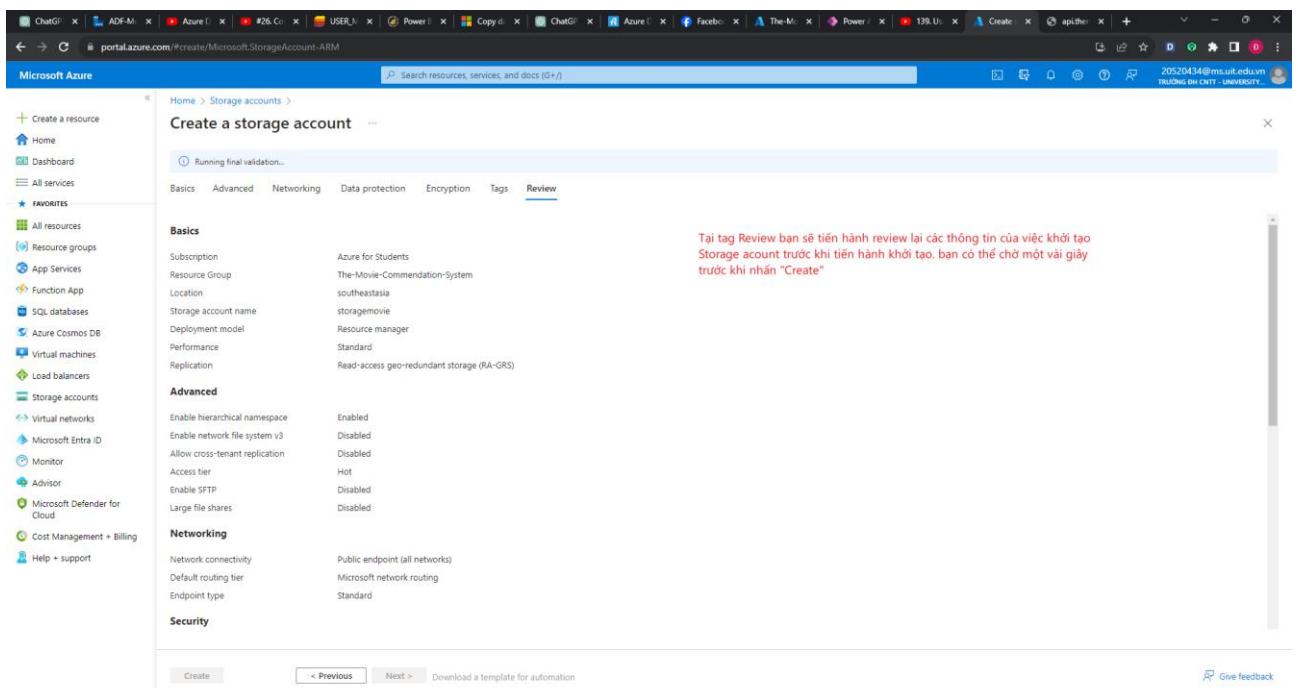
Hình 4.10. Khởi tạo các thông tin cơ bản của Storage tại Networking



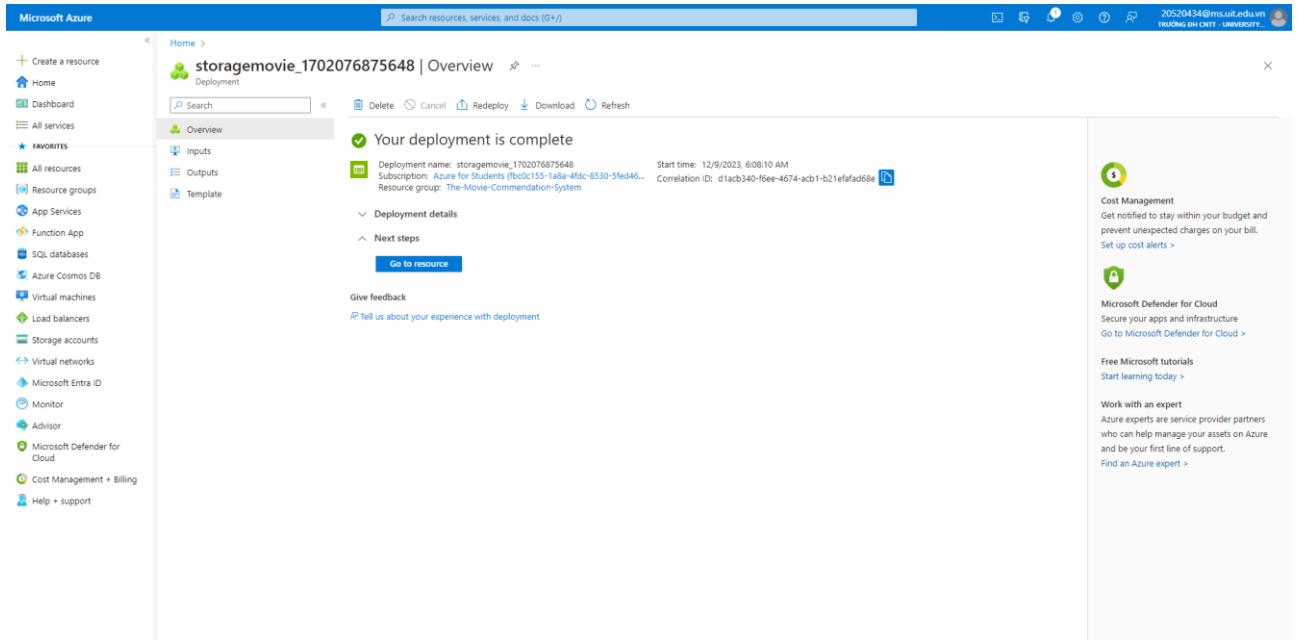
Hình 4.11. Khởi tạo các thông tin cơ bản của Storage tại Data protection



Hình 4.12. Khởi tạo các thông tin cơ bản của Storage tại Encryption and Tags

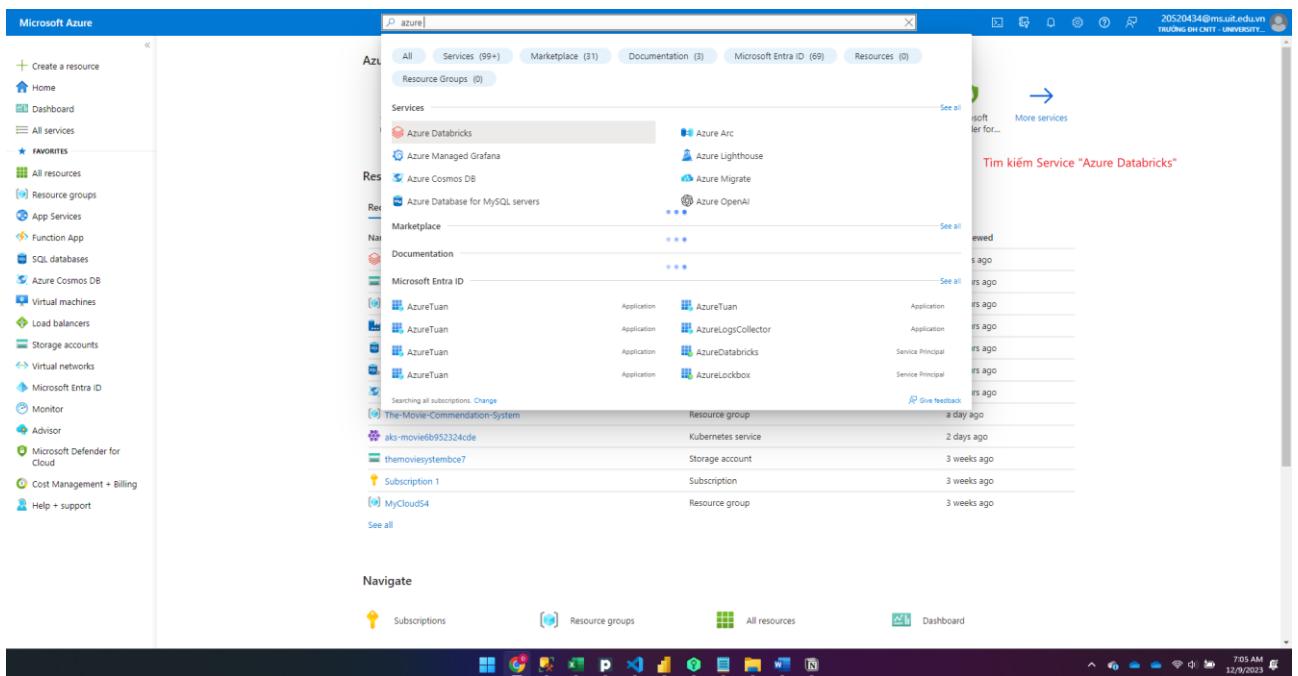


Hình 4.13. Review lại thông tin Storage Account trước khi Create tại tag Review



Hình 4.14. Khởi tạo Storage accounts thành công

4.3. Triển khai môi trường Azure Databricks



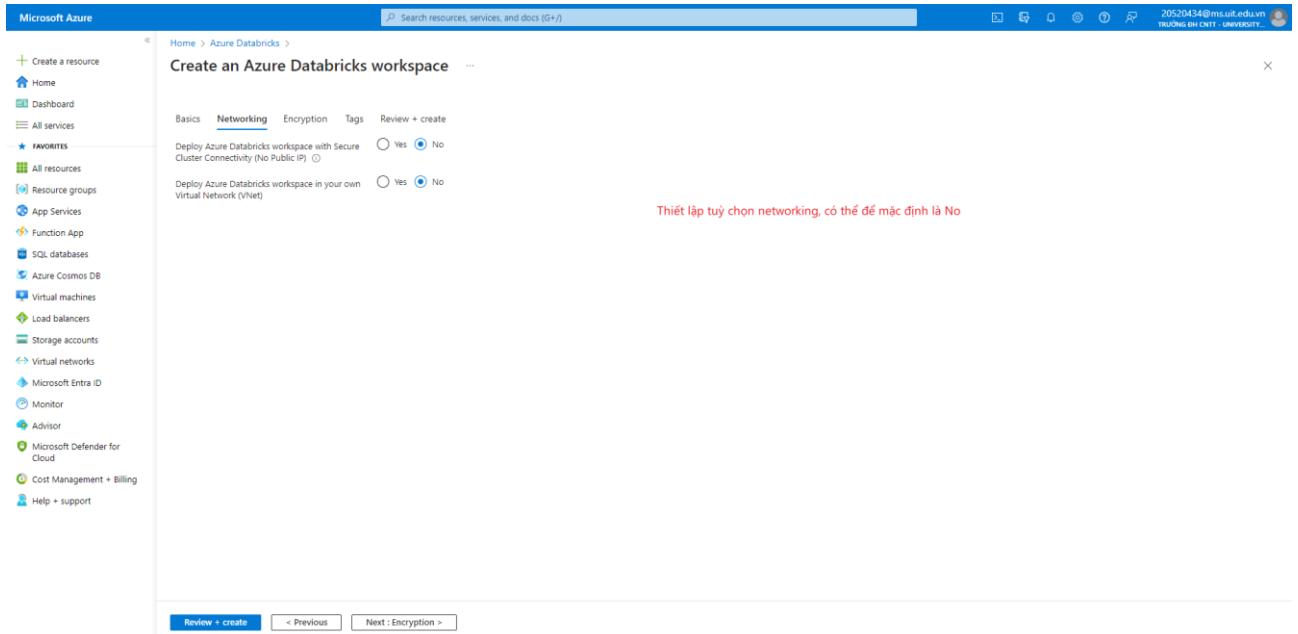
Hình 4.15. Tìm kiếm Services Azure Databricks

The screenshot shows the Microsoft Azure portal's search bar at the top with the query 'Azure Databricks'. Below it, the 'Create' button is highlighted with a red box. The main area shows a table with one row for 'The-Movie-System', which is an Azure Databricks Service located in East US with the subscription 'Azure for Students'. Navigation links like '< Previous', 'Page 1 of 1', and 'Next >' are at the bottom.

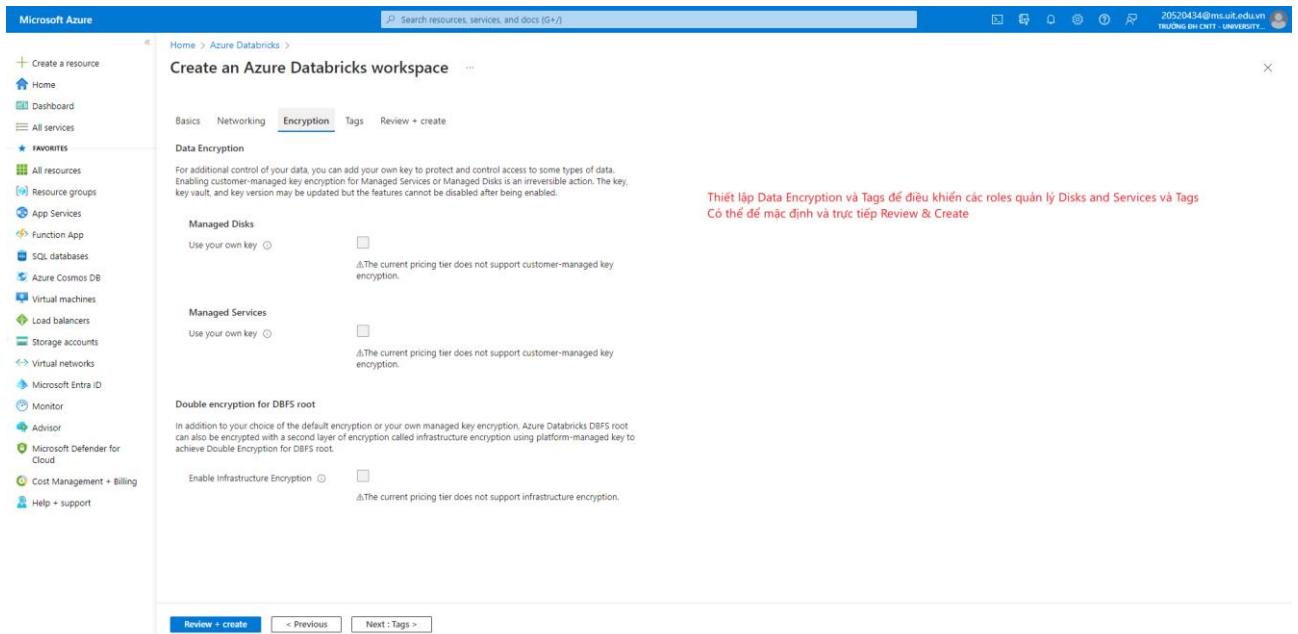
Hình 4.16. Tạo mới Azure Databricks

The screenshot shows the 'Create an Azure Databricks workspace' wizard. The 'Basics' tab is active. It includes fields for 'Subscription' (set to 'Azure for Students'), 'Resource group' (set to 'The-Movie-Commandment-System'), 'Workspace name' (set to 'System-Recommendation_Movie'), 'Region' (set to 'East US'), and 'Pricing Tier' (set to 'Trial (Premium - 14-Days Free DBUs)'). A red arrow points from the 'Subscription' dropdown to the 'Resource group' dropdown. Red annotations provide instructions: 'thực hiện điền tên các thông tin Subscription, và Resoucegroup như những lần khởi tạo trước' (Fill in the subscription and resource group names as in previous creations) and 'Điền thông tin workspace name và Region theo quy định và vị trí' (Fill in the workspace name and region according to regulations and location).

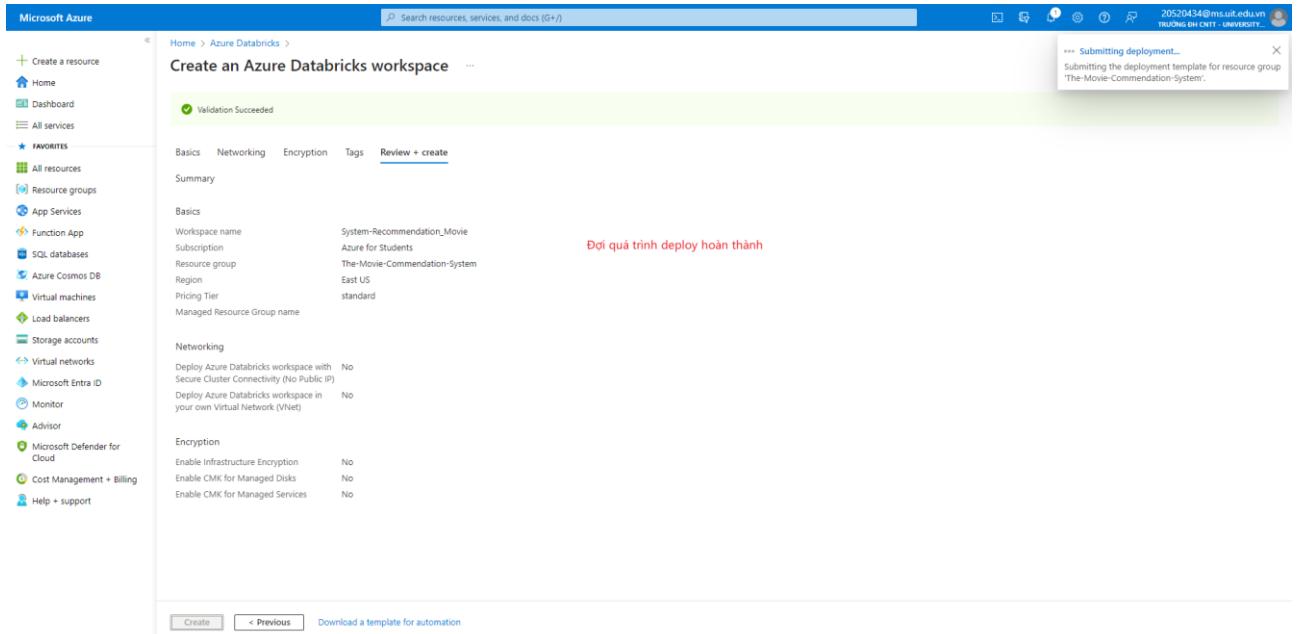
Hình 4.17. Thiết lập thông tin cơ bản tại Basics



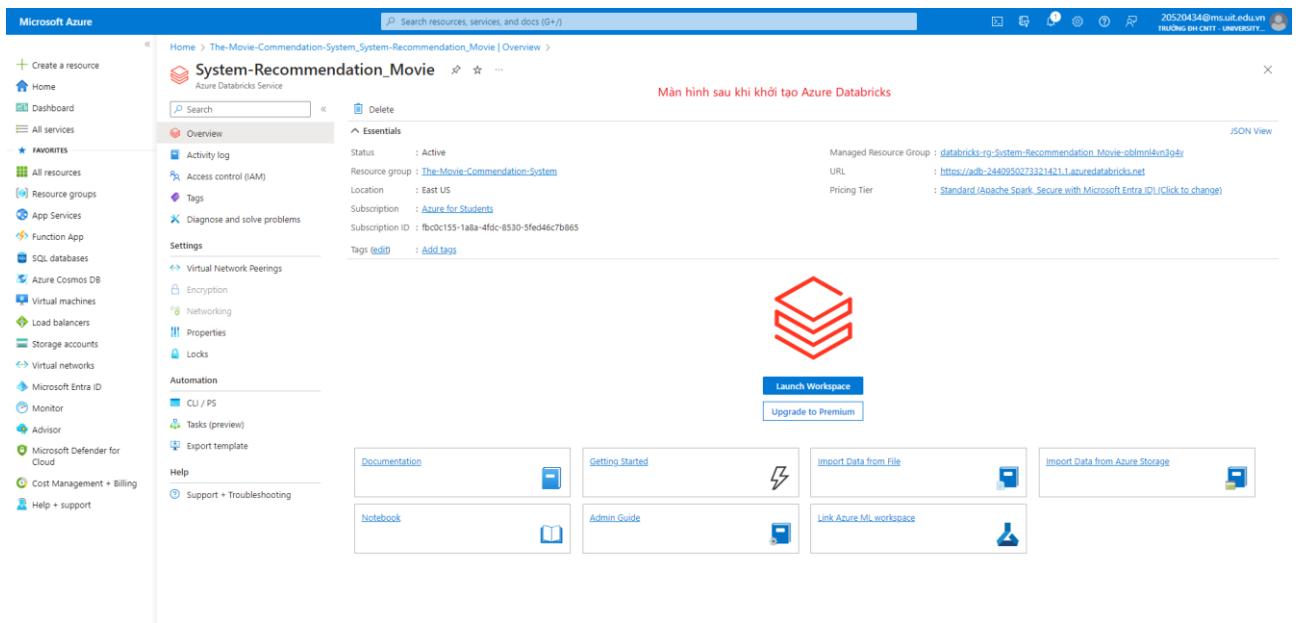
Hình 4.18. Thiết lập Networking để khởi tạo Databricks



Hình 4.19. Thiết lập các thông tin Ancrytion and Tags



Hình 4.20. Đợi kết quả Deploy



Hình 4.21. Màn hình sau khi khởi tạo Azure Databricks

Tiến hành thiết lập các tham số để tạo mới cluster

- + Performance:
 - Databricks Runtime Version: 14.2 (Includes Apache Spark 3.5.0 Scale 2.12)
 - Use Photo Acceleration sử dụng Photo Acceleration
 - Worker type: Sử dụng luồng Standard_DS3_v2 và tùy chọn autoscaling để tự động thiết lập số workers khi thực hiện
 - Advanced options: < có thể để mặc định>

Summary

2-8 Workers	28-112 GB Memory
8-32 Cores	1 Driver 14 GB Memory, 4 Cores
Runtime 14.2.x-scal2.12	Photon Standard_DS3_v2 4-14 DBU/h

Summary cấu hình cluster khởi tạo

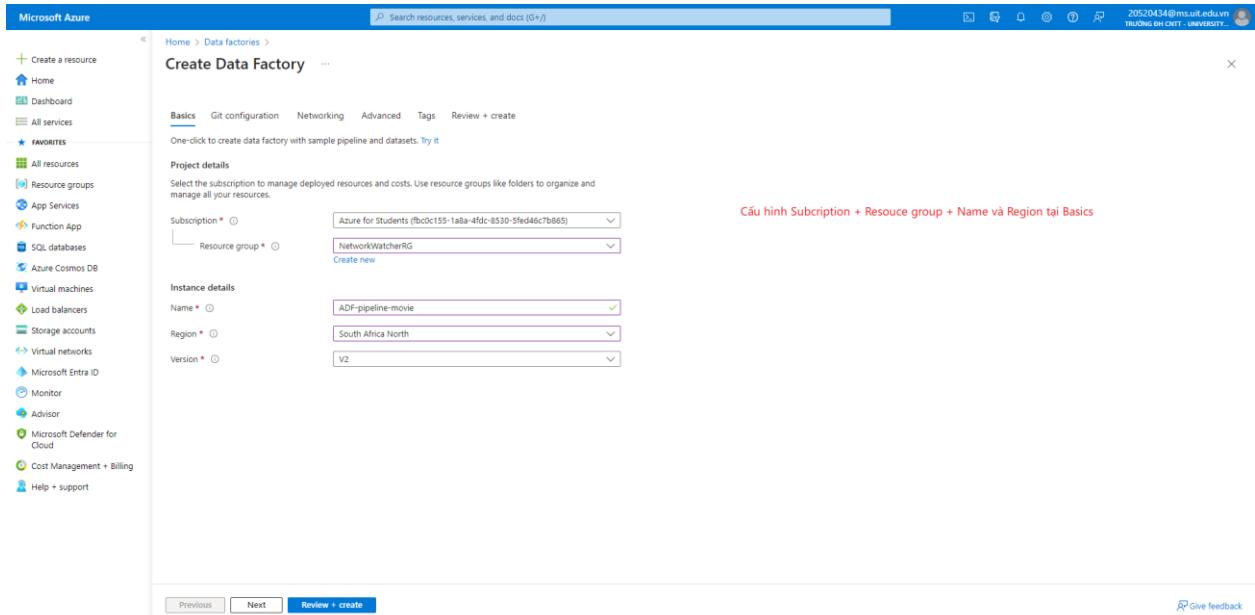
Hình 4.22. Tiến hành tạo mới Cluster với các thông số

4.4. Triển khai môi trường Azure Data Factory

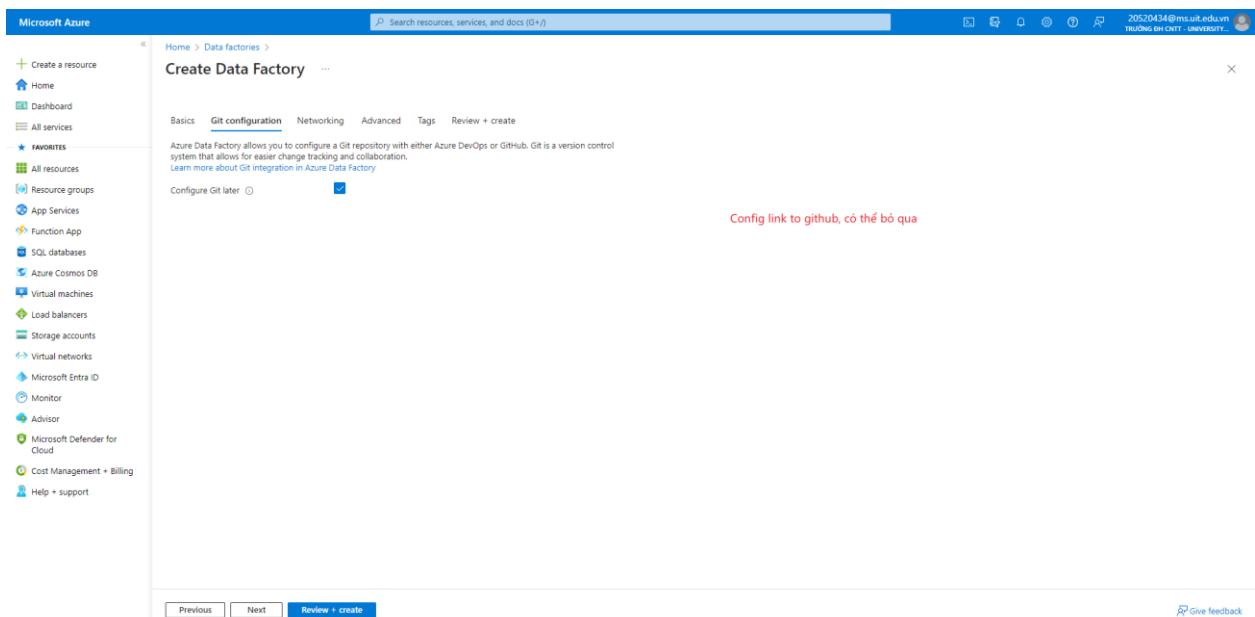
1. Tìm kiếm data factory

2. Chọn Services data factory

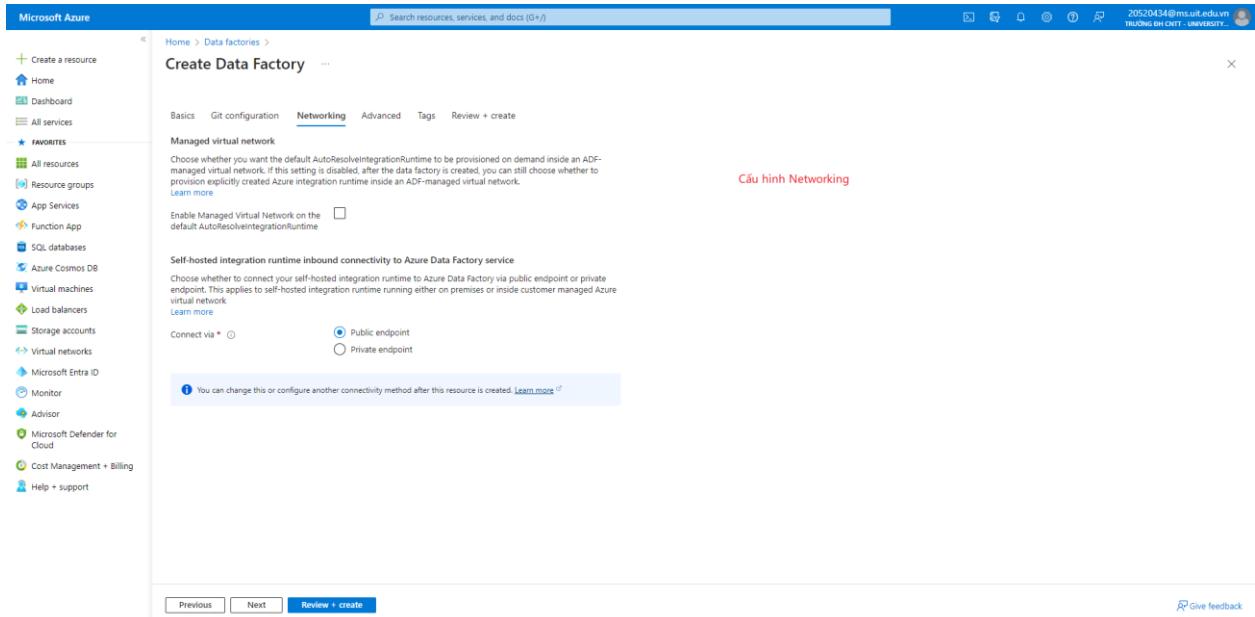
Hình 4.23. Tìm kiếm Data Factory



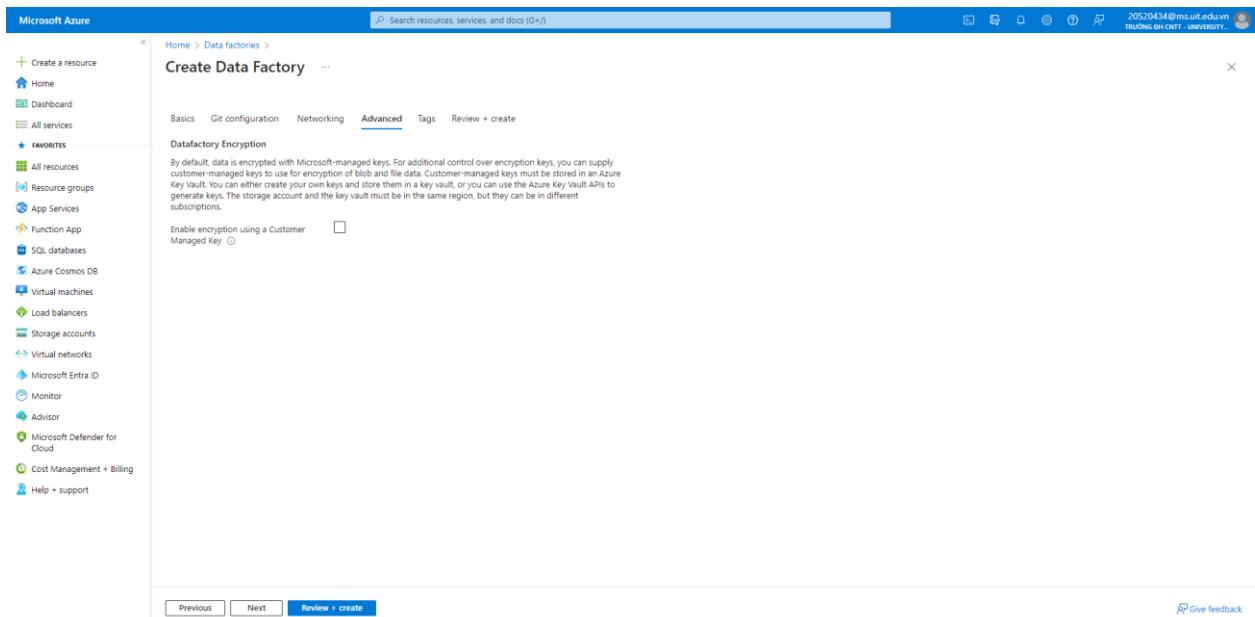
Hình 4.24. Cấu hình thông tin cơ bản tại Base



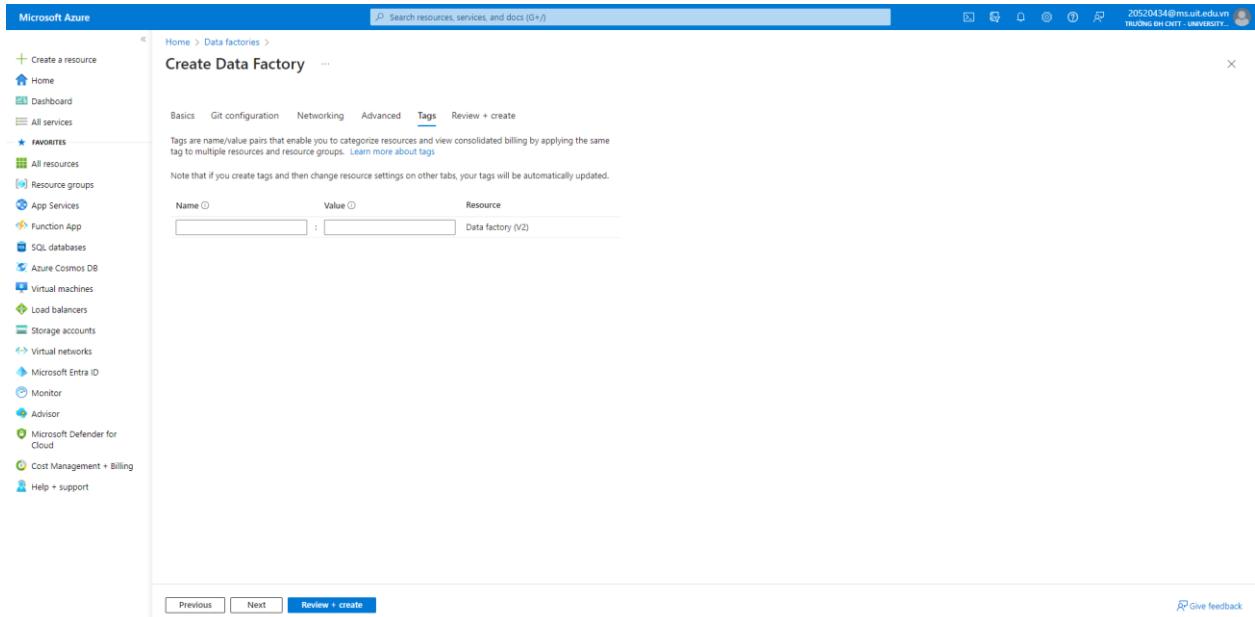
Hình 4.25. Cấu hình liên kết tới Github



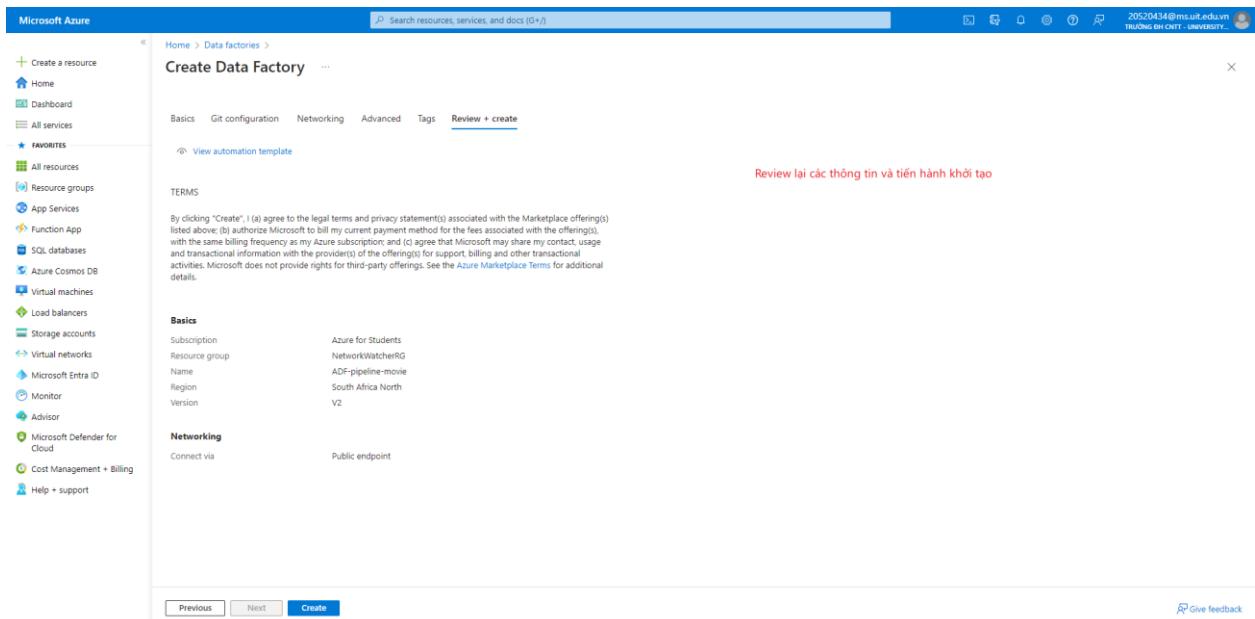
Hình 4.26. Cấu hình networking



Hình 4.27. Cấu hình advanced



Hình 4.28. Cấu hình Tags



Hình 4.29. Review lại thông tin và khởi tạo

The screenshot shows the Microsoft Azure portal interface. On the left, the sidebar includes 'Create a resource', 'Home', 'Dashboard', 'FAVORITES' (with 'All resources', 'Resource groups', 'App Services', 'Function App', 'SQL databases', 'Azure Cosmos DB', 'Virtual machines', 'Load balancers', 'Storage accounts', 'Virtual networks', 'Microsoft Entra ID', 'Monitor', 'Advisor', 'Microsoft Defender for Cloud', 'Cost Management + Billing', and 'Help + support'), and navigation links for 'Data factories' (selected), 'ADF-Movie-System', 'Overview', 'Delete', 'Essentials', 'Activity log', 'Access control (IAM)', 'Tags', 'Diagnose and solve problems', 'Properties', 'Locks', 'Getting started', 'Quick start', 'Monitoring', 'Alerts', 'Metrics', 'Diagnostic settings', 'Logs', 'Automation', 'CLI / PS', 'Tasks (preview)', 'Help', and 'Resource health'. The main content area displays the 'ADF-Movie-System' Data factory (V2) with details like Resource group (move), Status (Succeeded), Location (North Europe), Subscription (move), and Subscription ID. Below this is the 'Azure Data Factory Studio' interface with sections for 'Quick Starts', 'Tutorials', 'Template Gallery', and 'Training Modules', and monitoring dashboards for 'PipelineRuns' and 'ActivityRuns'.

Hình 4.30. Kết quả khởi tạo Data factories

The screenshot shows the 'Edit linked service' dialog in the Azure Data Factory studio. The left sidebar lists 'General', 'Connections' (selected), 'Source control', 'Author', 'Security', and 'Workflow orchestration manager'. The main area shows a table of 'Linked services' with three items: 'Connect_to_Blob' (Type: Azure Blob Storage), 'Connect_To_SQL' (Type: SQL server), and 'REST_API' (Type: REST). A red box highlights the 'Name' field for 'Connect_To_SQL'. The right side of the dialog shows configuration options for 'Connect via integration runtime' (AutoResolveIntegrationRuntime selected), 'Server name' (dev-the-movie-system.database.windows.net), 'Database name' (dev-rcm-movie), 'Authentication type' (SQL authentication), 'User name' (the-movie), and 'Password' (redacted). There are also tabs for 'Connection string' and 'Azure Key Vault', and sections for 'Description', 'Annotations', and 'Additional connection properties'.

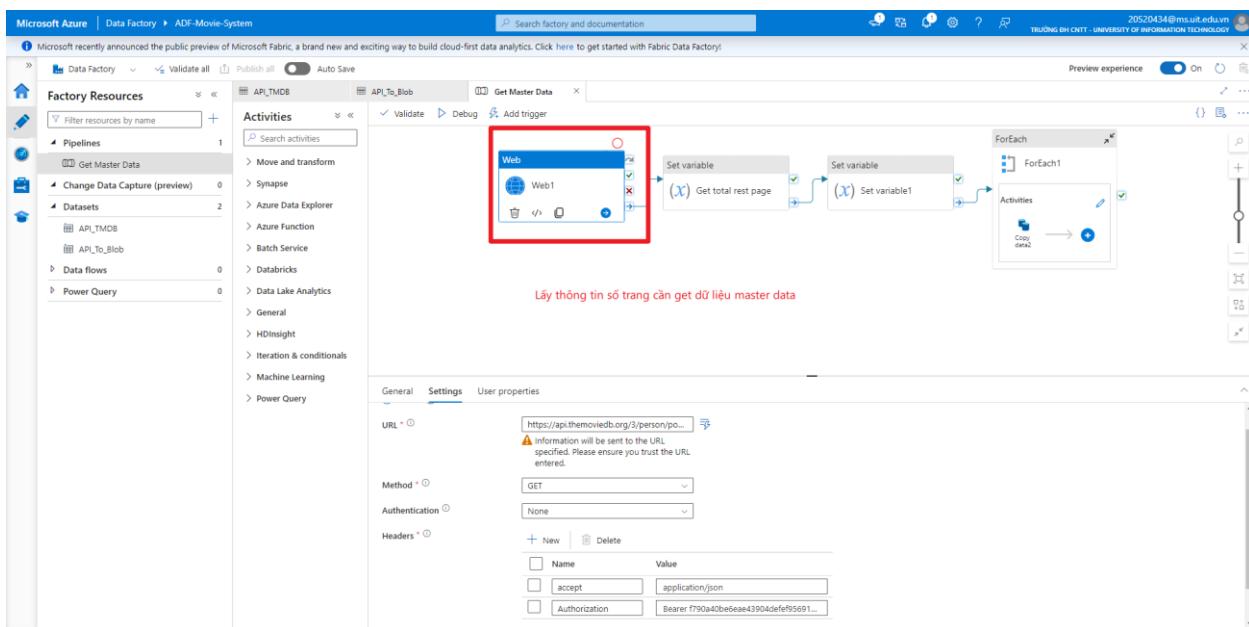
Hình 4.31. Khởi tạo link service tới SQL

The screenshot shows the Microsoft Azure Data Factory interface. On the left, a navigation sidebar lists various options like General, Connections, and Source control. In the main area, under 'Connections', 'Linked services' is selected. A table lists three existing linked services: 'Connect_to_Blob' (Azure Blob Storage), 'Connect_To_SQL' (SQL server), and 'REST_API' (REST). To the right, a detailed configuration pane is open for the 'REST_API' service. The 'Name' field is set to 'REST_API'. The 'Base URL' field contains the URL 'https://api.themoviedb.org/3/person/popular?api_key=f790a40be6ee43904defef95691aee3'. The 'Authentication type' is set to 'Anonymous'. Other settings include 'Server Certificate Validation' (Enable checked) and 'Auth headers'. Buttons for 'Save', 'Cancel', and 'Test connection' are at the bottom.

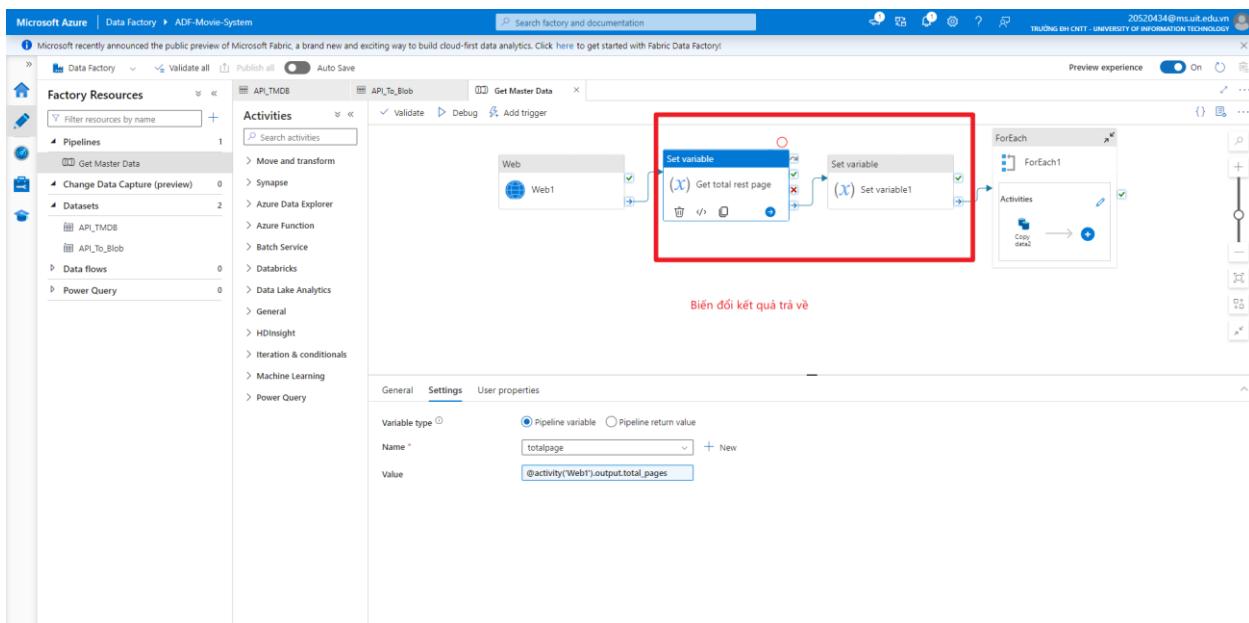
Hình 4.32. Khởi tạo link service tới API TMDB

This screenshot is similar to the previous one but shows the configuration for a different linked service. The 'Name' field in the configuration pane is now set to 'Connect_to_Blob'. The 'Base URL' field is empty. The 'Authentication type' is set to 'Account key'. Under 'Connection string', 'Azure Key Vault' is selected. The 'Storage account name' is 'masterdatamovie'. The 'Storage account key' field contains a masked value. Other fields like 'Partitioned DNS enabled', 'Endpoint suffix', and 'Additional connection properties' are shown below. Buttons for 'Apply', 'Cancel', and 'Test connection' are at the bottom.

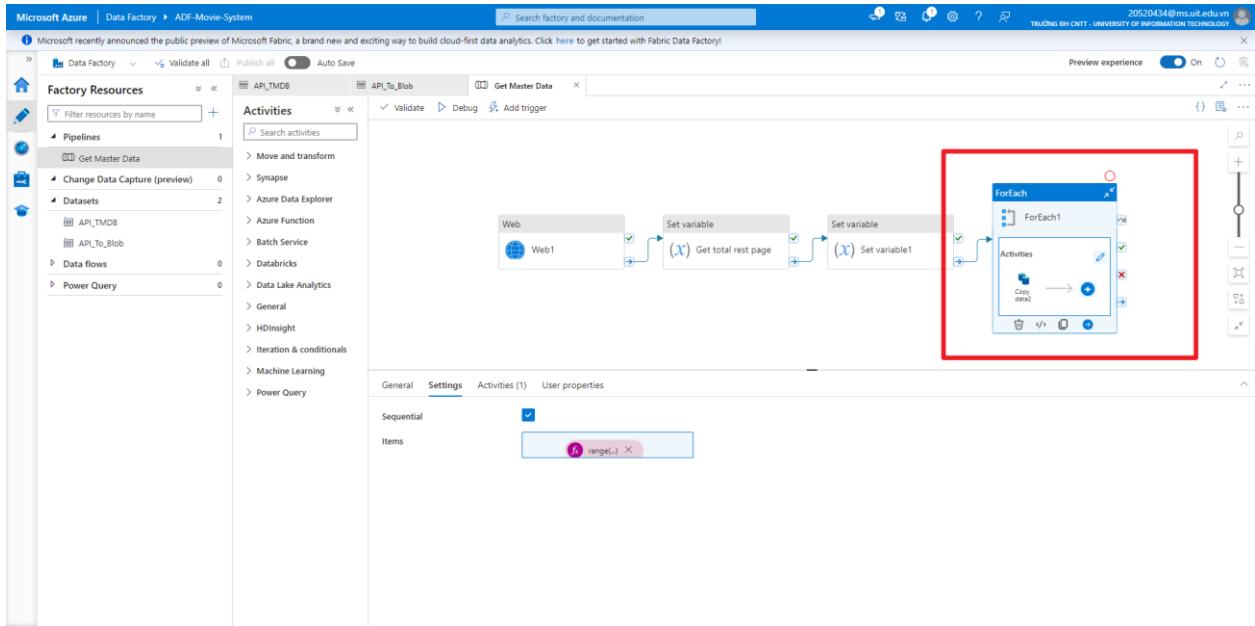
Hình 4.33. Khởi tạo link service tới Blob



Hình 4.34. Lấy thông tin dữ liệu tổng số trang



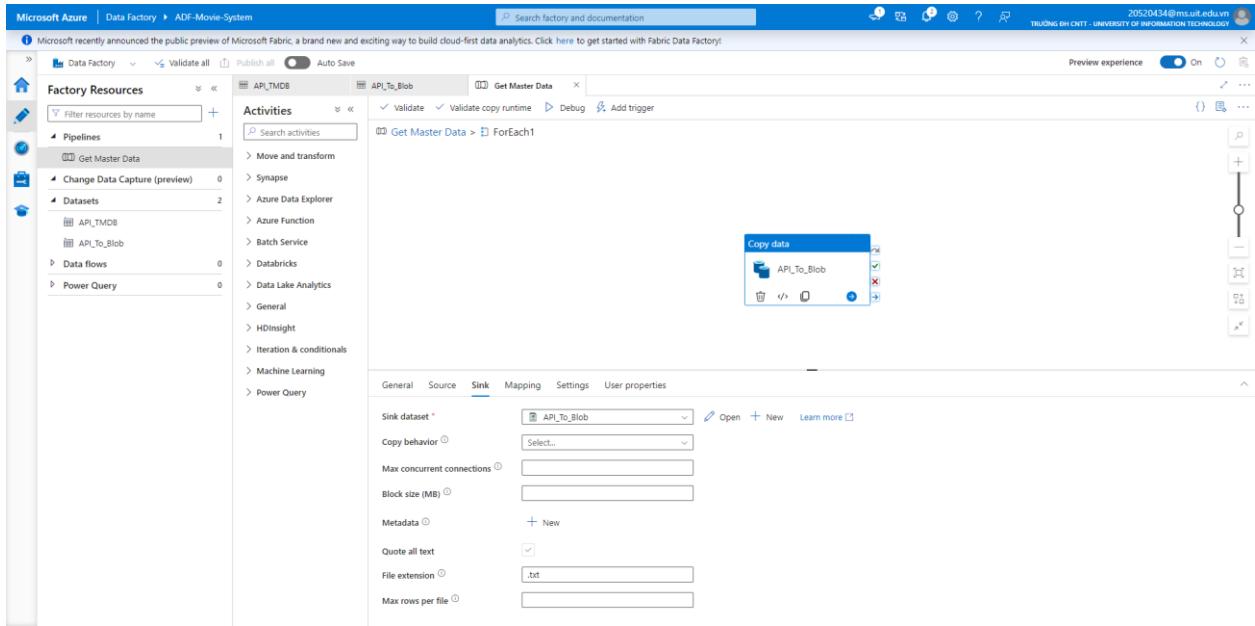
Hình 4.35. Biến đổi kết quả trả về dữ liệu để get total page



Hình 4.36. Vòng lặp Foreach duyệt qua lần lượt các trang

The screenshot shows the Microsoft Azure Data Factory Pipeline Designer interface. A 'Copy' activity is selected. The 'Source' tab is active, showing the configuration for the 'API_TMDB' dataset. It includes parameters such as 'API_KEY' (value: ?api_key=f790a40bbe6aa43904defe95d...) and 'page' (value: &page=@{item()}). Other tabs include 'Sink', 'Mapping', 'Settings', and 'User properties'.

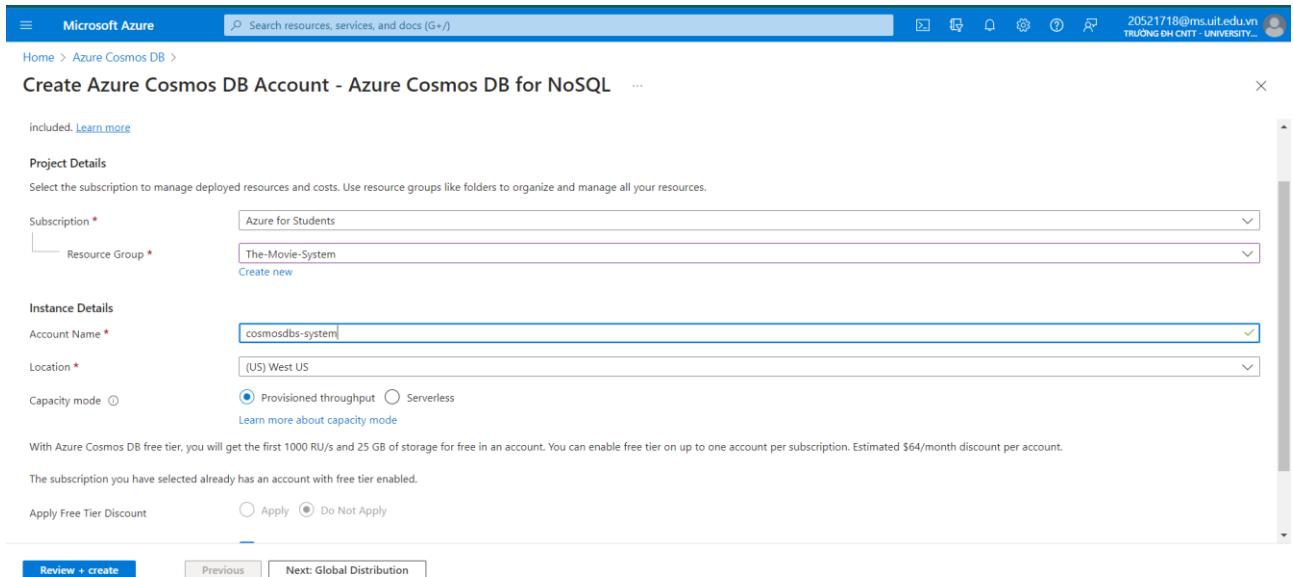
Hình 4.37. Cấu hình Source copy from TMDB



Hình 4.38. Cấu hình Sink to Blob Storage

4.5. Triển khai môi trường Azure Cosmos DB

Tạo resource Azure CosmosDB với cùng Resource Group với các resource khác, điền các thông tin cần thiết, các thông số mặc định, cuối cùng ấn “Review + create” để tạo resource.



Hình 4.39. Khởi tạo Resource Azure Cosmos DB for NoSQL

Xem lại list các resource Cosmos DB đã tạo.

The screenshot shows the Microsoft Azure portal interface for managing Azure Cosmos DB. At the top, there's a search bar and navigation links. Below it, the 'Azure Cosmos DB' section is visible, showing one record named 'cosmosdb-system'. The table includes columns for Name, Status, Subscription, Write region, and Read Region. The status is 'Online', and the subscription is 'Azure for Students'. Both regions listed are 'West US'. There are also filter and sorting options at the top of the table.

Hình 4.40. Xem các Database đã tạo

Thông tin tổng quát về Resource Cosmos DB vừa khởi tạo.

This screenshot shows the detailed view of the 'cosmosdb-system' database within the Azure Cosmos DB service. On the left, a sidebar lists various management options such as Overview, Activity log, Tags, Diagnose and solve problems, Cost Management, Quick start, Notifications, Data Explorer, and Settings. The main pane displays the 'Essentials' section, which includes the status (Online), resource group (The-Movie-System), subscription (Azure for Students), and specific throughput settings (1000 RU/s). It also shows the database's URI and free tier discount status. Below this, the 'Containers' section lists a single container named 'ctrnrcmmovie' with a throughput of 1000 RU/s. A 'JSON View' button is available on the right side of the essentials panel.

Hình 4.41. Thông tin về Database đã khởi tạo

Xem thông tin về Database và Collection (Container) sau khi đã triển khai đưa dữ liệu đã xử lý từ Azure DataBrick trong phần Data Explorer.

The screenshot shows the Microsoft Azure Data Explorer interface for the 'cosmosdb-system' database. The left sidebar includes options like 'Overview', 'Activity log', 'Access control (IAM)', 'Tags', 'Diagnose and solve problems', 'Cost Management', 'Quick start', 'Notifications', 'Data Explorer', and 'Settings'. The main pane displays a table with columns 'id', 'r_id', 'RCH_MOVIE_ID', and '_etag'. One row is selected, showing the JSON document content:

```

1   "RCH_MOVIE_ID": "[2393,5723,20186,26291,18082",
2   "id": "1183993",
3   "_rid": "w29JA2fkns48AAAAAAA==",
4   "_ts": 1682905273,
5   "_etag": "\\"97005273-0000-0700-0000-657061e50",
6   "_attachments": "attachments/",
7   "_ts": 17018053909
8
9

```

Hình 4.42. Thông tin về container trong DataExplorer

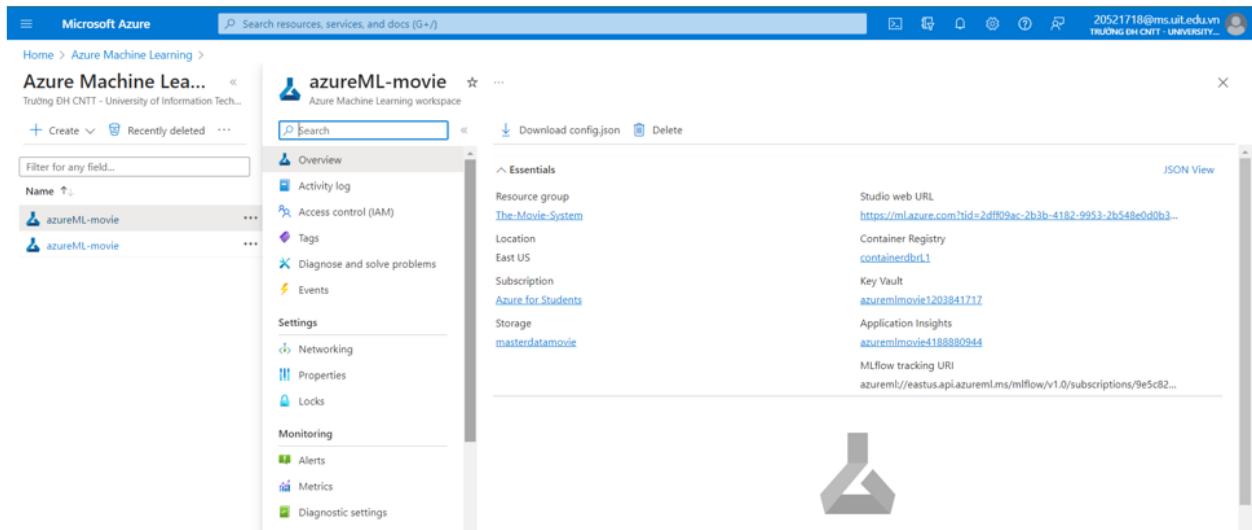
4.6. Triển khai môi trường Azure Machine Learning Service

Tạo resource Azure Machine Learning với cùng resource group với các service khác, Storage Account là account cùng với DB storage ban đầu, các thông số khác theo mặc định mà Azure khởi tạo, Án “Next” để xem và thao tác với các thông số nếu cần thiết, Án “Review + create” để tạo resource.

The screenshot shows the Azure portal interface for creating a new Azure Machine Learning workspace. The 'Resource details' section includes 'Subscription' (Azure for Students) and 'Resource group' (The-Movie-System). The 'Workspace details' section includes 'Name' (azureML-system), 'Region' (East US), 'Storage account' (masterdatamovie), 'Key vault' ((new) azuremlsystem9908600240), and 'Application insights' ((new) azuremlsystem1662225284). At the bottom, there are 'Review + create' and 'Next : Networking' buttons.

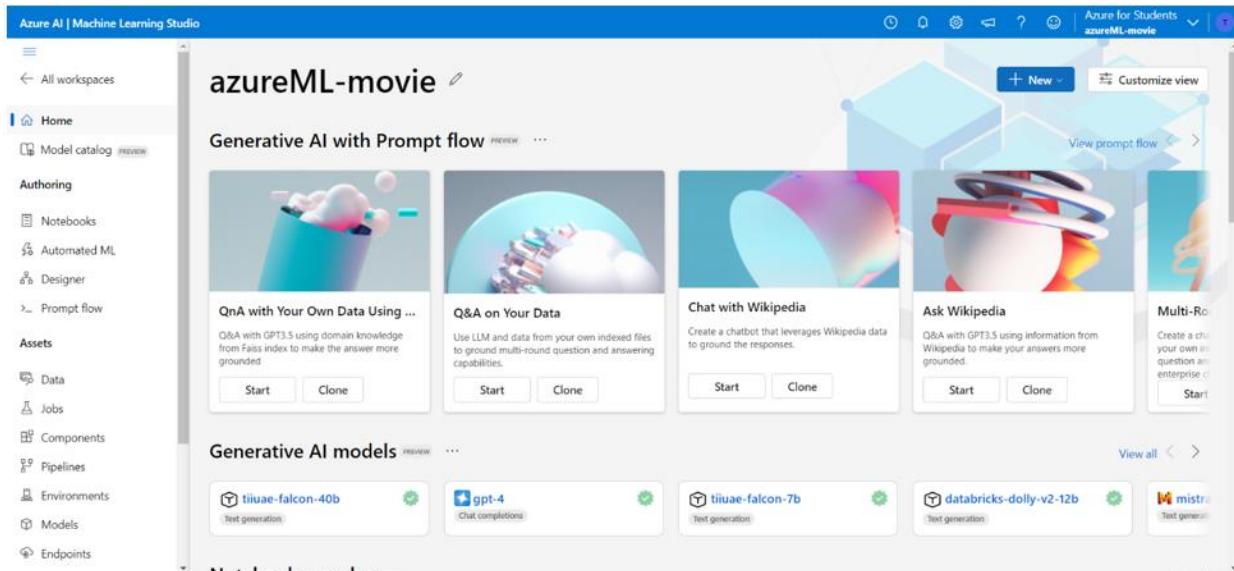
Hình 4.43. Khởi tạo Resource Azure ML

Thông tin về resource Azure Machine Learning vừa khởi tạo tại phần Overview.



Hình 4.44. Thông tin tổng quát Resource Azure ML đã tạo

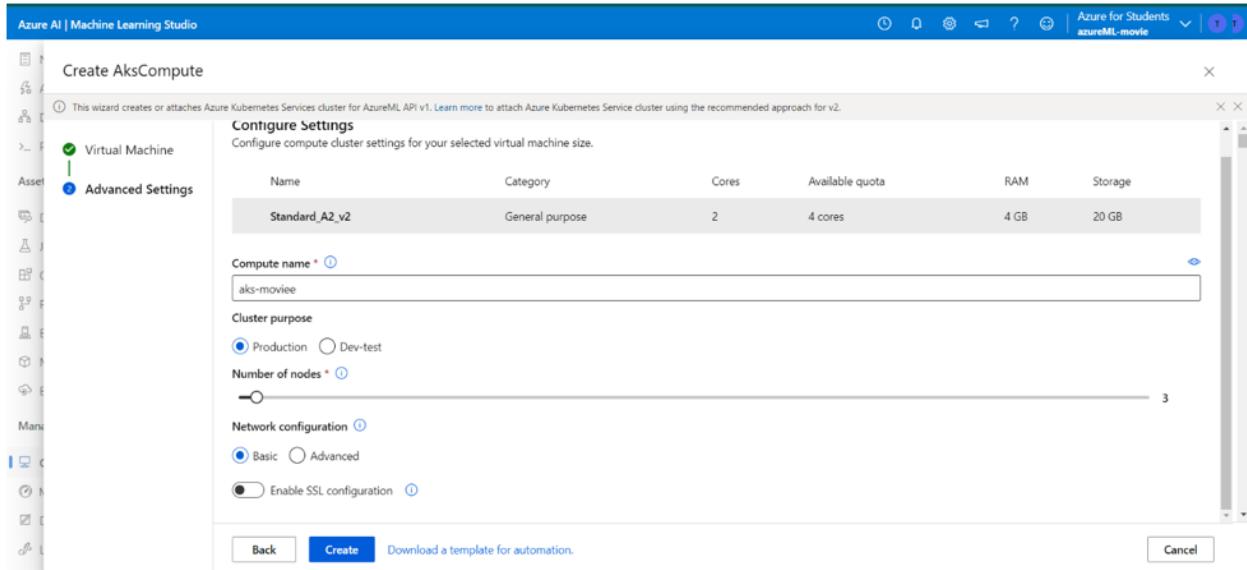
Truy cập Azure ML Studio



Hình 4.45. Giao diện Azure ML studio

4.7. Triển khai môi trường Azure Kubernetes Service

Tạo Compute trong Azure ML với AksCompute, chọn cấu hình Compute thích hợp, chọn vào “Create new” để khởi tạo mới 1 Compute, chọn “Next” để sang bước kế tiếp.



Hình 4.46. Khởi tạo AksCompute trong Azure ML studio

Ở bước này, Ta tiến hành đặt tên cho Compute và tinh chỉnh các thông số về “Cluster purpose” và “Number of nodes” (mặc định là 3), Án chọn “Create” để deploy Compute.

Kubernetes services được tự động tạo khi Compute được tạo thành công, dung để quản lí các thông tin liên quan đến AksCompute.

Essentials	Properties
Resource group the-movie-system	Kubernetes version 1.27.7
Status Succeeded (Running)	API server address aks-movie6b952324cde-9v4pkj3.hcp.eastus.azmk8s.io
Location East US	Network type (plugin) Kubenet
Subscription Azure for Students	Node pools 1 node pool
Subscription ID 9e5c82be-8447-4074-aa9d-86d3d9e8a6de	
Tags (edit) Add tags	

Hình 4.47. Cài đặt các thông số cho AksCompute

Thông tin về Endpoint tương ứng với Compute đã tạo sao khi đã deploy Webservice thành công.

Name	Description	Quota type	Created on	Created by	Updated on	Compute type
aks-movie-rcm	-	-	Dec 8, 2023 8:33 AM	Nhị Tôn Nữ Thảo	Dec 8, 2023 8:33 AM	AksCompute

Hình 4.48. Danh sách các Endpoint trong Azure ML studio

Endpoint attributes	Properties
Service ID aks-movie-rcm	hasInferenceSchema True
Description -	hasHttps False
Deployment state Healthy	authEnabled True
Compute type AksCompute	
Created by Nhị Tôn Nữ Thảo	
Model ID aks-movie-reco.mml:1	
Created on Dec 8, 2023 8:33 AM	
Last updated on Dec 8, 2023 8:33 AM	

Hình 4.49. Thông tin chi tiết AksCompute - I

Azure AI | Machine Learning Studio

Trường DH CNTT - University of Information Technology > azureML-movie > Compute > aks-movie

aks-movie ☆

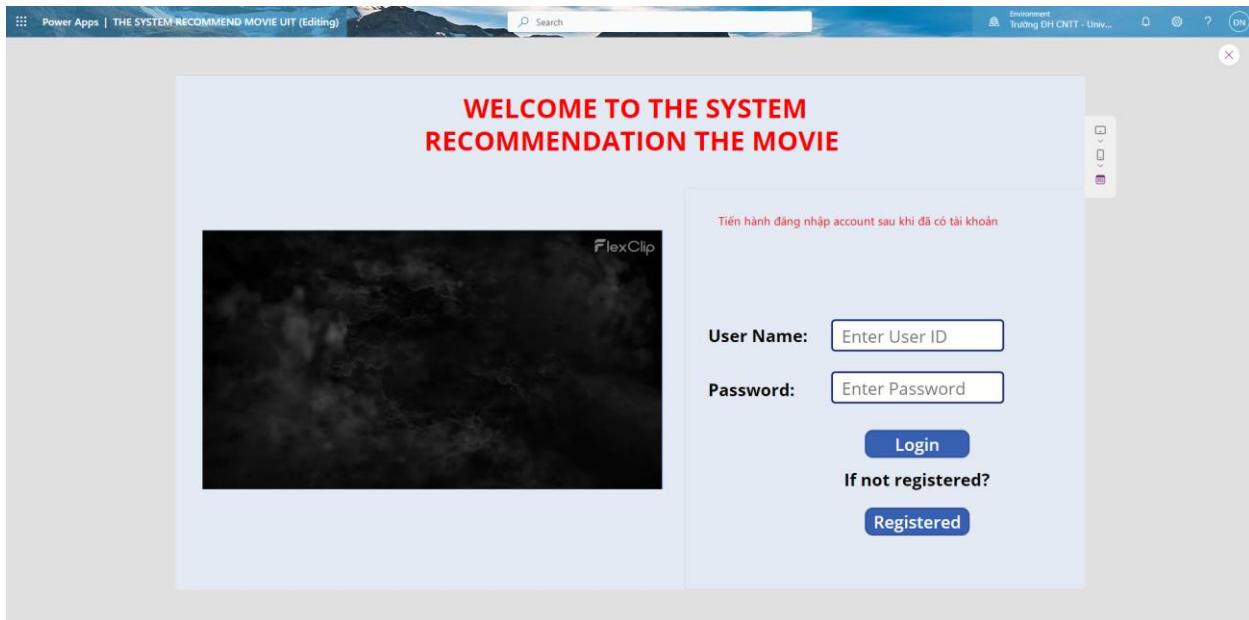
Details Monitoring (preview)

Attributes

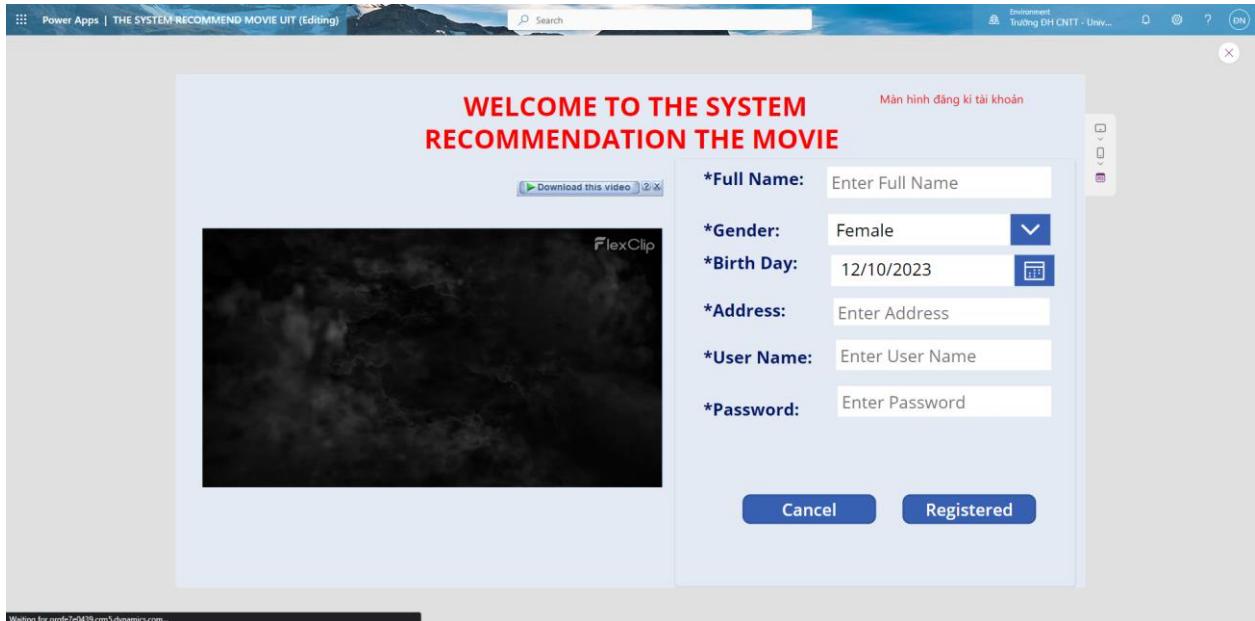
- Compute name: aks-movie
- Cluster name: aks-movie6b952324cde
- Compute type: AksCompute
- Subscription ID: 9e5c82be-8447-4074-aa9d-86d3d9e8a6de
- Resource group: The-Movie-System
- Workspace: azureML-movie
- Region: eastus
- Allocation state: Succeeded

Hình 4.50. Thông tin chi tiết AksCompute - 2

4.8. Triển khai ứng dụng trên Power Apps



Hình 4.51. Màn hình đăng nhập vào hệ thống

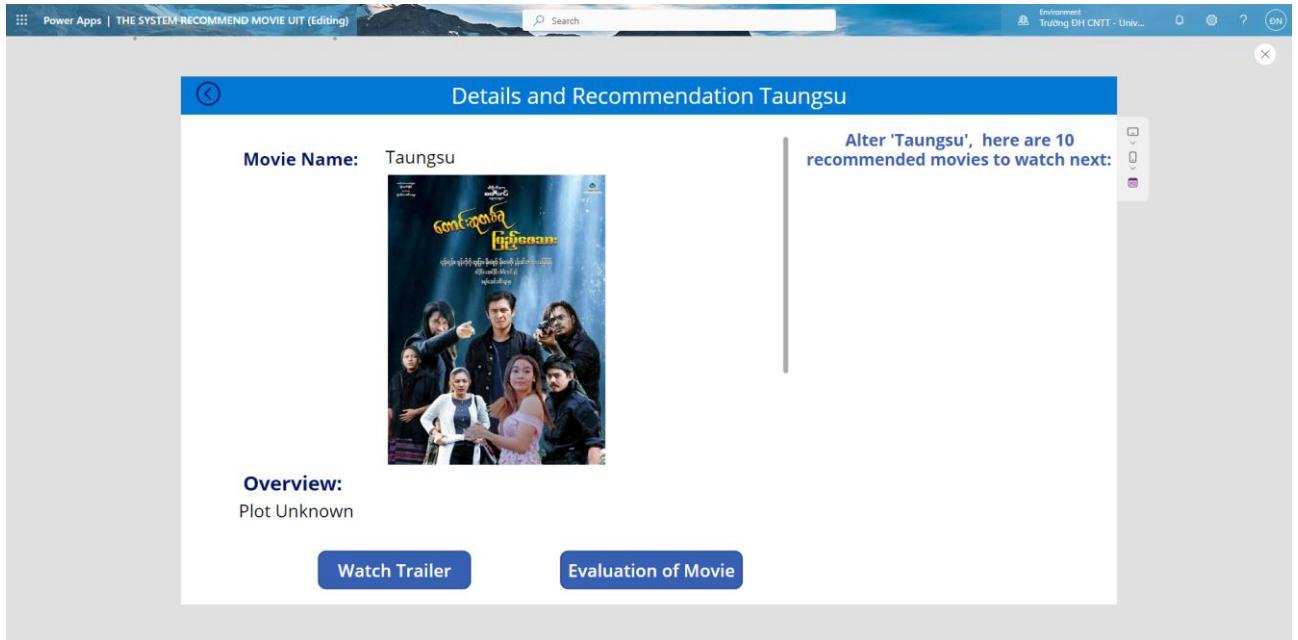


Hình 4.52. Màn hình đăng ký tài khoản

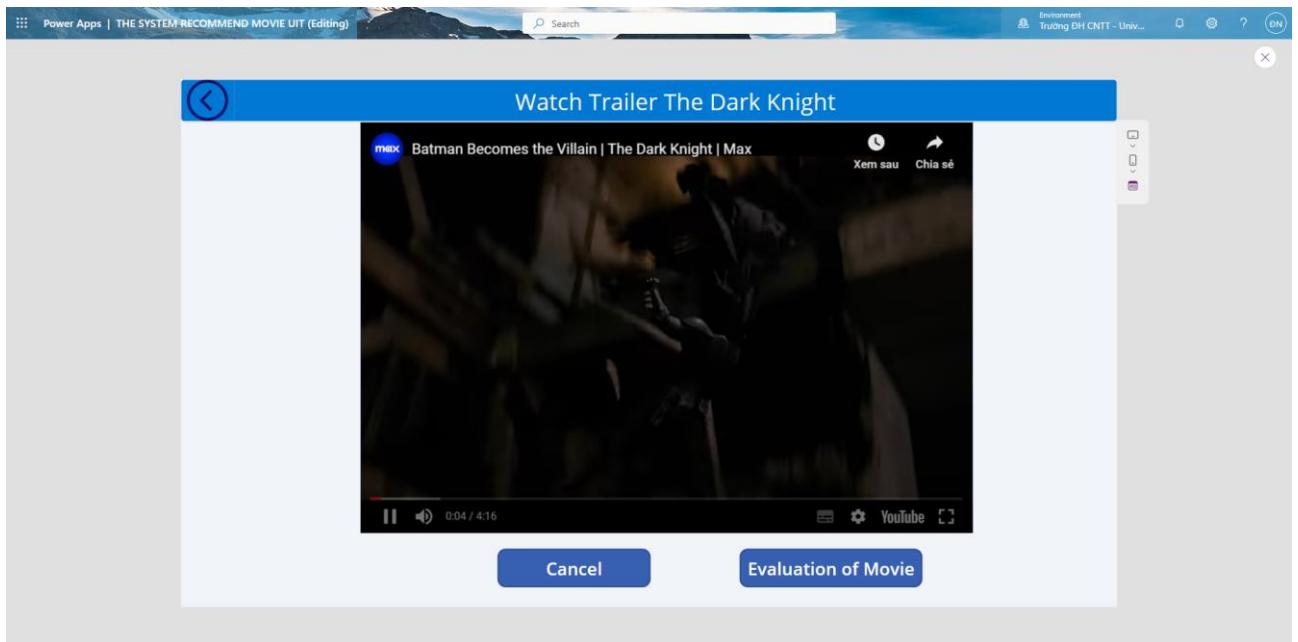
The System Movie Recommendation

Movie Title	Plot Summary
Chamán	The old Shaman makes his last trip to the Amazon jungle in search of a special beeswax for arrows that he learned to make
Taungsu	Plot Unknown
Day 13	A horror feature that tells the story of Ezz El-Din, who returns from abroad after many years, looking for his family, and while
As Long As We Both Shall Live	An unhappy young woman from an abusive family is married off to a fearsome and chilly army commander. But the two learn
No Direction Home	After missing the last train after their best friend's wedding, Ko, Baill, and Vacumaca are stranded off at a small train station in

Hình 4.53. Màn hình tìm kiếm bộ phim và đánh giá



Hình 4.54. Màn hình xem thông tin chi tiết phim và phim recommendation



Hình 4.55. Màn hình xem trailer Phim

Evaluation of Movie Dark Heaven

Full Name*: Hồ Hữu Thuyết Address*: Lâm Đồng, Việt Nam

Year Old: 18

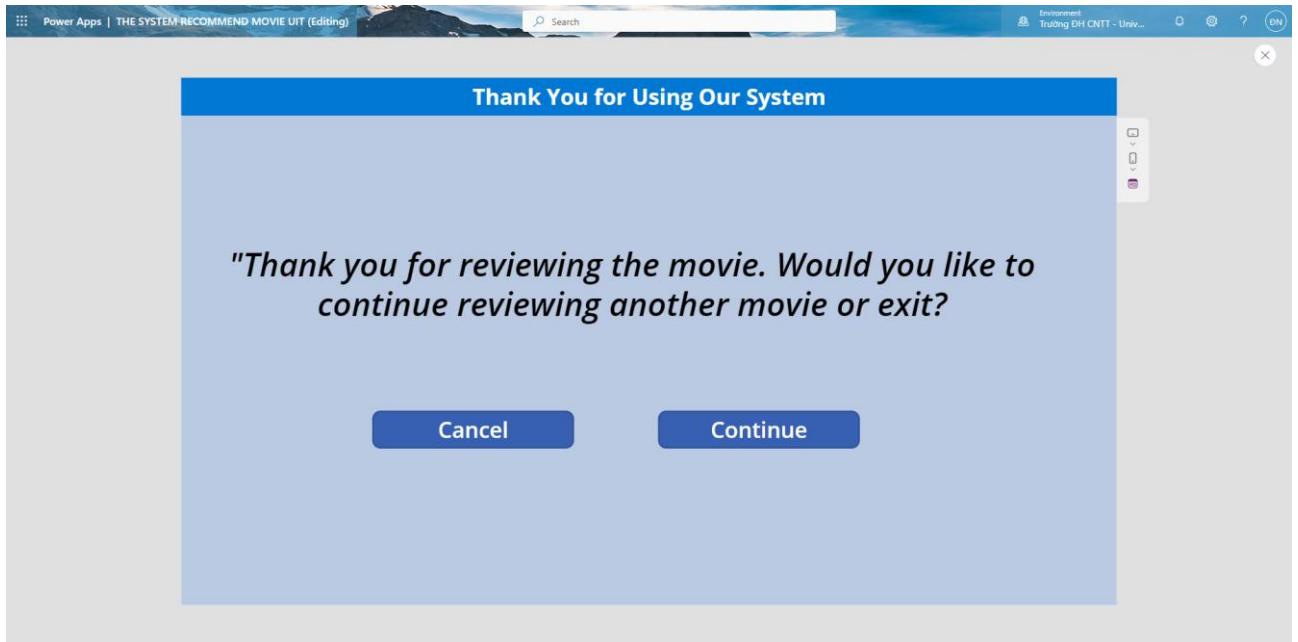
Did you enjoy the movie? True False

Date Watched: 12/8/2023

Rating: ★ ★ ★ ★ ★ ★ ★ ★ ★

Comment:

Hình 4.56. Màn hình đánh giá phim



Hình 4.57. Màn hình xuất thông báo

4.9. Triển khai toàn bộ hệ thống

Cài đặt file Azure-cosmos-spark vào Libraris của Spark Cluster để tương tác với CosmosDB từ Azure DataBricks.

The screenshot shows the Databricks UI for a cluster named "The Movie Rcm's Cluster". The left sidebar is collapsed, and the main area is titled "Libraries". A single JAR file is listed:

Status	Name	Type	Source
-	azure_cosmos_spark_3_4_2_12_4_24_0.jar	JAR	dbfs:/FileStore/jars/bada8e94_2d3d_4c79_8a98_6f26142f7f58-azure_cosmos_spark_3_4_...

Buttons for "Uninstall" and "Install new" are visible at the top right of the library list.

Hình 4.58. Cài đặt file JAR trong Libraries của Cluster

CHƯƠNG 5. KẾT LUẬN

5.1. Kết quả đạt được

Hệ thống đã được triển khai thành công trên nền tảng điện toán đám mây Microsoft Azure, bao gồm các dịch vụ Azure Blob Storage, Azure Data Factory, Azure Databricks, Azure Machine Learning Service, Azure Cosmos DB và Azure Kubernetes Service.

Xây dựng được hệ thống đề xuất phim dựa trên sở thích và đánh giá của người dùng.

Áp dụng được 2 thuật toán đề xuất phim phổ biến là Alternating Least Square (ALS) và Surprise Singular Value Decomposition (SVD).

Sử dụng PowerApps để thể hiện hệ thống đề xuất phim, cung cấp khả năng mở rộng và dễ dàng tích hợp với các hệ thống khác trong môi trường Azure, tạo điều kiện cho việc phát triển và cải tiến liên tục của hệ thống.

5.2. Hạn chế

Hệ thống phụ thuộc vào các dịch vụ điện toán đám mây của Microsoft Azure, do đó chi phí triển khai và vận hành hệ thống có thể cao.

Dữ liệu đầu vào cho hệ thống đề nghị phim của nhóm bao gồm dữ liệu về phim, dữ liệu về người dùng và dữ liệu về các tương tác giữa người dùng và phim. Dữ liệu này còn có một số hạn chế như không đầy đủ, không chính xác và không nhất quán. Điều này gây ảnh hưởng đến độ chính xác của hệ thống đề nghị.

Giá trị RMSE của thuật toán ALS và SVD còn quá cao do dữ liệu đầu vào chưa được tốt và còn hạn chế về huấn luyện mô hình.

PowerApps của nhóm để lấy và hiển thị thông tin phim cũng như các phim đề nghị còn có nhiều hạn chế như giao diện chưa đẹp, thông tin của phim chưa đầy đủ và còn thiếu một số các chức năng khác.

5.3. Hướng phát triển

Có thể sử dụng các dịch vụ điện toán đám mây của các nhà cung cấp khác để giảm chi phí triển khai và vận hành hệ thống.

Cải thiện chất lượng dữ liệu đầu vào: Nhóm cần cải thiện chất lượng dữ liệu đầu vào bằng cách thu thập dữ liệu đầy đủ, chính xác và nhất quán. Điều này có thể được thực hiện bằng cách sử dụng các phương pháp thu thập dữ liệu tiên tiến, chẳng hạn như học máy từ dữ liệu lớn.

Nhóm có thể sử dụng các kỹ thuật bổ sung để cải thiện độ chính xác của các thuật toán ALS và SVD như sử dụng các kỹ thuật xử lý dữ liệu tiên tiến và các thuật toán tối ưu hóa mới để giúp cải thiện hiệu quả của thuật toán ALS và SVD.

Nghiên cứu các mô hình đề nghị mới: Nhóm cần tiếp tục nghiên cứu các mô hình đề nghị mới để cải thiện độ chính xác của hệ thống đề nghị. Các mô hình đề nghị mới cần có khả năng xử lý các dữ liệu đa dạng và phức tạp.

Power Apps: Cải thiện PowerApps về giao diện, chức năng và hiển thị đầy đủ thông tin các bộ phim cũng như hệ thống gợi ý phim.

Tăng cường hiệu suất: Nhóm cần tìm cách tăng cường hiệu suất của hệ thống đề nghị bằng cách sử dụng các kỹ thuật như phân tích song song, nén dữ liệu và tối ưu hóa thuật toán.

BẢNG PHÂN CÔNG CÔNG VIỆC

	Đạt	Ánh	Cường	Nhi
Lựa chọn đề tài	x	x	x	x
Tổng quan đề tài	x			
Khảo sát, tìm hiểu dataset	x	x	x	x
Thu thập dữ liệu	x		x	
Giới thiệu hệ thống Azure		x		x
Quy trình ra quyết định	x	x		x
Triển khai model		x		x
Các bước triển khai trên Azure	x		x	
Xây dựng ứng dụng	x		x	
Slide	x	x	x	x
Viết báo cáo	x	x	x	x
Chỉnh sửa format báo cáo		x		

Bảng 1. Bảng phân công công việc

TÀI LIỆU THAM KHẢO

- [1] fptcloud, "Microft Azure," [Online]. Available: <https://microsoft.fptcloud.com/microsoft-azure/>. [Accessed 11 2023].
- [2] pacisoft, "Giới thiệu về Azure | Microsoft Azure," pacisoft.vn, [Online]. Available: <https://www.pacisoft.vn/gioi-thieu-ve-microsoft-azure/>. [Accessed 11 2023].
- [3] V. T. D. Guong, "Tìm hiểu về Azure Blob Storage trong Microsoft Azure," viblo, [Online]. Available: <https://viblo.asia/p/tim-hieu-ve-azure-blob-storage-trong-microsoft-azure-3P0IPDBPlox>. [Accessed 11 2023].
- [4] H. N. DOanh, "Windows Azure Blob Storage," viblo, [Online]. Available: <https://viblo.asia/p/windows-azure-blob-storage-ZDEeLXDoeJb>.
- [5] medium, "What is a Blob," medium, [Online]. Available: <https://medium.com/geekculture/what-is-a-blob-83e65f590694>.
- [6] datascientest, "Azure Data Factory: what is it and what is it for?," datascientest, [Online]. Available: <https://datascientest.com/en/azure-data-factory-what-is-it-and-what-is-it-for>. [Accessed 11 2023].
- [7] microsoft, "What is Azure Databricks?," learn.microsoft.com, [Online]. Available: <https://learn.microsoft.com/en-us/azure/databricks/introduction/>.
- [8] analyticsvidhya, "Azure Databricks: Key Features, Use Cases and Benefits," analyticsvidhya, [Online]. Available: <https://www.analyticsvidhya.com/blog/2023/02/azure-databricks-a-comprehensive-guide/#:~:text=Introduction,learning%20tasks%20easily>. [Accessed 11 2023].

- [9] funix, "microsoft azure machine learning va automl hop ly hoa quy trinh may hoc," funix, [Online]. Available: <https://funix.edu.vn/chia-se-kien-thuc/microsoft-azure-machine-learning-va-automl-hop-ly-hoa-quy-trinh-may-hoc/>. [Accessed 11 2023].
- [10] T. Sáng, "ppt_ALS," scribd, [Online]. Available: <https://www.scribd.com/presentation/543477582/ppt-ALS>. [Accessed 12 2023].
- [11] KevinLiao159, "movie_recommender/movie_recommendation_using_ALS.ipynb," github, [Online]. Available: https://github.com/KevinLiao159/MyDataSciencePortfolio/blob/master/movie_recommender/movie_recommendation_using_ALS.ipynb. [Accessed 12 2023].
- [12] H. D. Quân, "Công nghệ Matrix Factorization cho Hệ thống gợi ý," viblo, [Online]. Available: <https://viblo.asia/p/cong-nghe-matrix-factorization-cho-he-thong-goi-y-naQZRJe0Zvx>. [Accessed 12 2023].
- [13] T. Trần, "Xây dựng hệ thống Recommendation với model ALS," linkedin, [Online]. Available: <https://www.linkedin.com/pulse/x%C3%A2y-d%E1%BB%B1ng-h%E1%BB%87-th%E1%BB%91ng-recommendation-v%E1%BB%9Bi-model-als-t%C3%A2m-tr%E1%BA%A7n>. [Accessed 12 2023].
- [14] miguelgfierro, "examples/02_model_collaborative_filtering/surprise_svd_deep_dive.ipynb," github, [Online]. Available: https://github.com/recommenders-team/recommenders/blob/main/examples/02_model_collaborative_filtering/surprise_svd_deep_dive.ipynb. [Accessed 12 2023].
- [15] surprise, "Matrix Factorization-based algorithms," surprise, [Online]. Available: https://surprise.readthedocs.io/en/stable/matrix_factorization.html. [Accessed 12 2023].

- [16] microsoft, "Azure Cosmos DB – Unified AI Database," microsoft, [Online]. Available: <https://learn.microsoft.com/en-us/azure/cosmos-db/introduction>. [Accessed 11 2023].
- [17] microsoft, "What is Azure Container Instances?," microsoft, [Online]. Available: <https://learn.microsoft.com/en-us/azure/container-instances/container-instances-overview>. [Accessed 11 2023].
- [18] microsoft, "What is Azure Kubernetes Service?," microsoft, [Online]. Available: <https://learn.microsoft.com/en-us/azure/aks/intro-kubernetes>. [Accessed 11 2023].
- [19] microsoft, "What is Power Apps?," microsoft, [Online]. Available: <https://learn.microsoft.com/en-us/power-apps/powerapps-overview>. [Accessed 11 2023].
- [20] microsoft, "Copy file from SharePoint Online," microsoft, [Online]. Available: <https://learn.microsoft.com/en-us/azure/data-factory/connector-sharepoint-online-list?tabs=data-factory#copy-file-from-sharepoint-online>. [Accessed 12 2023].
- [21] miguelgfierro, "recommend/SETUP.md," github, [Online]. Available: <https://github.com/recommenders-team/recommenders/blob/main/SETUP.md#repository-installation>. [Accessed 12 2023].
- [22] H. Đinh, "Xây dựng Collaborative Filtering RS [Recommender System cơ bản - Phần 3]," viblo, [Online]. Available: <https://viblo.asia/p/xay-dung-collaborative-filtering-rs-recommender-system-co-ban-phan-3-Az45bMqolxY>. [Accessed 12 2023].

- [23] microsoft, "Build a real-time recommendation API on Azure," microsoft, [Online]. Available: <https://learn.microsoft.com/en-us/azure/architecture/ai-ml/architecture/real-time-recommendation>. [Accessed 10 2023].
- [24] miguelgfierro, "recommenders-team/recommenders," github, [Online]. Available: <https://github.com/recommenders-team/recommenders>. [Accessed 10 2023].