

**PHÂN TÍCH VÀ TRỰC QUAN DỮ LIỆU TÌNH HÌNH KINH
DOANH CỦA CÔNG TY**

Azure Data Factory, Databrick, SQL Database & PowerBI

DANH MỤC BẢNG

Bảng 1. So sánh tốc độ xử lý giữa máy chủ truyền thống và hệ thống cloud	4
--	---

DANH MỤC HÌNH ẢNH

Hình 1. Tổng quan về Công ty Lending Club	1
Hình 2. Đưa ra quyết định dựa trên dữ liệu của doanh nghiệp.....	2
Hình 3. Kiến trúc tổng quan của hệ thống	5
Hình 4. Quá trình ETL. Nguồn: Extract, transform, load (ETL) - Azure Architecture Center Microsoft Learn.....	6
Hình 5. Mô hình Star schema. Nguồn: Mô hình hóa dữ liệu: Star Schema, Bảng DIM và bảng FACT (tomanhhoang.com).....	7
Hình 6. Sự kết hợp giữa quá trình ETL và mô hình Star schema. Nguồn: Data Modelling: Techniques, Importance and Implementation by Venkatakrishnan Medium	10
Hình 7. Tổng quan kĩ thuật xử lý dữ liệu được sử dụng	11
Hình 8. Thành phần của Blob Storage. Nguồn: k21academy.com	13
Hình 9. Ba tầng truy cập của Blob Storage. Nguồn: linkedin.com	14
Hình 10. Azure Data Factory - Usecase	15
Hình 11. Các Linked Services ví dụ. Nguồn: cathrinewilhelmsen.net.....	17
Hình 12. Minh họa giữa Activities và Pipelines. Nguồn: Cagthrine Wilhelmsen (hi@cathrinew.net)	18
Hình 13. Mapping Data Flows. Nguồn: clearpeaks.com	18
Hình 14. Thành phần cơ bản trong Databricks. Nguồn: faun.pub	19
Hình 15. Hai mô hình triển khai trong Azure SQL Databasse. Nguồn: cloudvietnam18.com.....	21
Hình 16. Ưu và nhược điểm việc sử dụng Azure SQL Database làm Data warehouse	23
Hình 17. Cơ chế chế độ truy vấn - Direct Query trên PowerBI	23
Hình 18. Cấu hình linked service - dataset với nguồn dữ liệu Azure Blob Storage	25

Hình 19. Tổng quan quy trình Transform trong Azure Data Factory	25
Hình 20. Chi tiết code xem qua file notebook data_pre_processing.ipynb.....	26
Hình 21. Tổ chức dữ liệu thành các bảng Dimensions (Mappping Data flow)	26
Hình 22. Tạo bảng Fact từ các bảng Dim và dữ liệu ban đầu (Mapping Data flow)	27
Hình 23. Cấu hình Linked service đến các bảng trong Azure SQL Database	27
Hình 24. Các datasets kết nối với các Linked services	27
Hình 25. Mô hình Star schema trong Data warehouse sau quá trình ETL.....	28
Hình 26. Tốc độ thực thi của các thành phần trong Data pipeline	28
Hình 27. Thực hiện đo độ trễ giữa 2 server ở 2 region khác nhau	29
Hình 28. Trang tổng quan.....	30
Hình 29. Trang filter 1: Đặc trưng về các khoản vay	30
Hình 30. Trang filter 2: Đặc trưng về người vay.....	31
Hình 31. Trang chi tiết về các thông số của công ty	31
Hình 32. Xuất bản lên PowerBI Service	32
Hình 33. Giao diện trên PowerBI service.....	32
Hình 34. Kết nối với mô hình dữ liệu sử dụng Direct Query	33
Hình 35. Cấu hình điều kiện cho các vai trò người dùng trong giao diện desktop	34
Hình 36. Chỉ định vai trò cho các tài khoản người dùng truy cập	34
Hình 37. Một số giao diện dashboard trên mobile	35
Hình 38. Thành phần điều hướng giữa các trang report.....	35
Hình 39. Slicer cho report với các thành phần như slider, combo-box, map.....	35
Hình 40. Các loại Phân vùng cho nguồn dữ liệu. Nguồn: learn.microsoft.com.....	36
Hình 41. Trong tab Resource groups, chọn Create để tạo mới một Resource group.....	37
Hình 42. Nhập các thông tin về tên, khu vực và tùy chọn gói tại tab Basics.....	37
Hình 43. Tại tab Review + create, sau khi Azure review các thông tin vừa nhập thì chọn Create để hoàn tất quá trình tạo Resource group.....	38
Hình 44. Chọn Create Resources để tạo những Resource cần trong Resource group	38
Hình 45. Tìm kiếm và chọn Storage account trong Marketplace.....	39

Hình 46. Sau khi Azure review lại thông tin vừa nhập ở các tab thì chọn Create để hoàn tất tạo Storage Account.....	39
Hình 47. Trong phần Data storage, chọn Containers và nhấn + Container để tạo mới.....	40
Hình 48. Tạo mới hai Container là "inputraw" và "inputclean" và upload các file csv vào	40
Hình 49. Vào Marketplace chọn Create SQL Database	41
Hình 50. Nhập các thông tin và chọn Create new ở mục Server.....	41
Hình 51. Nhập các thông tin và nhấn Create để tạo mới Server	42
Hình 52. Sau đó, khi hoàn thành các tab thì chọn Create để tạo mới SQL Database	42
Hình 53. Tiếp đó đợi Azure deploy thành công	43
Hình 54. Tiếp tục chọn Configure ở mục Configure access để cấu hình truy cập cho SQL	43
Hình 55. Cập nhật lại cấu hình như trên.....	44
Hình 56. Vào lại file SQL Database loan_data, chọn Query editor và Stored Procedures để upload truy vấn	44
Hình 57. Sau đó chọn Run để chạy các câu truy vấn	45
Hình 58. Quay lại Marketplace tìm Databricks chọn Create để tạo mới	45
Hình 59. Nhập thông tin các tab, chọn Review + create và đợi Azure review xong thì chon Create để tạo mới	46
Hình 60. Deploy thành công	46
Hình 61. Tại đây chọn Launch Workspace để sử dụng.....	47
Hình 62. Sau khi đăng nhập, tạo mới Compute bằng cách chọn Create Compute	47
Hình 63. Setup lại Cluster và chọn Create	48
Hình 64. Tại menu chọn New và Notebook để tạo mới	48
Hình 65. Chọn File và thực hiện Import notebook từ máy	49
Hình 66. Vào lại Storage Account lấy key và paste vào notebook	49
Hình 67. Vào lại Marketplace để tạo mới Data Factory.....	50
Hình 68. Nhập thông tin các tab và thực hiện Review + Create	50

Hình 69. Ở giao diện Data Factory, chuyển tới mục Connections, chọn Linked Services và Create Linked Services để kết nối tới các dịch vụ cần thiết	51
Hình 70. Đầu tiên thiết lập dịch vụ Blob Storage như hình	51
Hình 71. Tiếp tục thiết lập dịch vụ SQL Database và nhập thông tin vào	52
Hình 72. Tiếp tục thiết lập dịch vụ Databricks	52
Hình 73. Chọn lại tab Author trong menu và Import from pipeline template từ máy	53
Hình 74. Sau khi import, cấu hình pipeline như trên hình và chọn Use this teamplate	53
Hình 75. Sau đó, tại activity Transform data, qua tab Settings chọn Notebook path đúng	54
Hình 76. Sau khi kiểm tra qua các activity còn lại, thực hiện Publish all và chọn Publish	54
Hình 77. Sau khi chọn Add trigger để chạy Pipeline, chờ cho Run thành công	55
Hình 78. Để kiểm tra xem dữ liệu đã được copy vào bảng trong database, quay lại SQL Database "loan_data", Chọn Query editor (preview) trong phần Settings để chạy thử câu truy vấn và trả kết quả như trên là đã xem như thành công	55
Hình 79. Data model trong PowerBI sau khi tạo các measure cần thiết cho các report (sử dụng hàm DAX)	56
Hình 80. Cấu hình RLS - Role: Long-term Joint Loan Analyst	56
Hình 81. Cấu hình RLS - Role: Mature Loan Analyst	57
Hình 82. Cấu hình RLS - Role: Regional Loan Processor	57
Hình 83. Thực hiện kết nối với Azure SQL DB, nhập đúng tên Server và chọn DirectQuery mode để tạo live connection	57
Hình 84. Sau khi load các bảng cần thiết thì ta có giao diện dashboard như trên, đầu tiên là report Overview	58
Hình 85. Report thứ 2: Loan Characteristics	58
Hình 86. Report thứ 3: Borrower Characteristics	59
Hình 87. Report 4: Detail Statistics	59
Hình 88. Sau khi save thì thực hiện Publish lên PowerBI service	60
Hình 89. Kiểm tra report trên PowerBI service đã upload lên chưa và thực hiện tương tác với report	60

Hình 90. Tương tác thử trên PowerBI service với các report, ví dụ report 2: thực hiện tương tác filter và trả kết quả theo Slicers.....	61
Hình 91. Tiếp theo là kiểm tra cập nhật dữ liệu tức thời, đầu tiên vào Azure SQL Database để cập nhật dữ liệu, đăng nhập và chọn Query editor để thực hiện	61
Hình 92. Thực hiện truy vấn thêm 2 ngày vào năm 2021 và 2022	62
Hình 93. Sau khi Refresh lại trang để dashboard cập nhật, ta có thể thấy thay đổi ở filter đã tăng range từ 2017-2020 lên năm 2017-2022	62
Hình 94. Tiếp theo đến phần thiết lập quyền bảo mật, chọn vào Manage roles trong tab Modeling và setup cho role Long-term Joint Loan Analyst như hình	63
Hình 95. Tạo role Manager với toàn quyền quản lý (không có filter ở các bảng)	63
Hình 96. Tạo role Mature Loan Analyst thiết lập quyền như hình	64
Hình 97. Role Regional Loan Processor có thiết lập quyền như hình	64
Hình 98. Thực hiện cấp quyền cho user, truy cập vào setting workspace, chọn Security của Semantic Model như hình	65
Hình 99. Test với role Mature Loan Analyst cấp cho user Lê Gia Kiệt và nhấn Add	65
Hình 100. Ta có thể thấy khi đăng nhập user Lê Gia Kiệt đã được cấp role Mature Loan Analyst thì chỉ có thể xem và tương tác trên các giới hạn được cấp (grade là A,B,C)	66
Hình 101. Thực hiện dashboard responsive trên thiết bị điện thoại ở các report, ví dụ Overview	66
Hình 102. Thực hiện tối ưu trong Data Flows.....	67
Hình 103. Tối ưu activity transformissuedate, numberduplicatecolumn1, numberduplicatecolumn2, numberduplicatecolumn3, filterduplicate1, filterduplicate2, filterduplicate3, sinkborrower, sinkloan, sinktime của import_dim_table : Bật phân vùng Set partitioning trong tab Optimize, điều chỉnh sang Round Robin	67
Hình 104. Tối ưu joinborrower, joinloan, jointime, sinkfact của import_fact_table : Bật phân vùng Set partitioning trong tab Optimize, điều chỉnh sang Round Robin	68
Hình 105. Sau đó Publish all	68
Hình 106. Add trigger để chạy pipeline	69
Hình 107. Kết quả trả về.....	69

MỤC LỤC

I. GIỚI THIỆU BÀI TOÁN	1
1. Tổng quan bài toán.....	1
2. Tổng quan về bộ dữ liệu.....	3
3. So sánh tốc độ xử lý máy chủ truyền thống và hệ thống cloud	4
4. Kiến trúc dự án	5
II. CƠ SỞ LÝ THUYẾT	6
1. Định dạng lưu trữ	6
1.1. Quá trình ETL (Extract, Transform, Load)	6
1.2. Mô hình hoá dữ liệu – Star schema	7
1.3. Thiết kế kho dữ liệu với sự kết hợp ETL và mô hình Star schema.....	8
2. Kỹ thuật xử lý dữ liệu.....	11
2.1. Đọc dữ liệu	11
2.2. Xử lý giá trị bị thiếu	11
2.3. Chuyển đổi kiểu dữ liệu.....	11
2.4. Làm sạch dữ liệu	11
2.5. Tính toán các cột mới.....	12
2.6. Tổng hợp dữ liệu	12
2.7. Mô hình hoá dữ liệu (Star schema)	12
2.8. Kết hợp sử dụng Azure Databricks và Mapping data flow cho việc biến đổi dữ liệu	12
3. Các thành phần.....	13
3.1. Sử dụng Azure Blob Storage để lưu trữ dữ liệu nguồn	13
3.1.1. Giới thiệu	13

3.1.2. Các thành phần liên quan	13
3.1.3. Đặc điểm	14
3.2. Điều phối bằng Azure Data Factory.....	14
3.2.1. Extract	16
3.2.2. Transform.....	16
3.2.3. Load	16
3.2.4. Linked Services	16
3.2.5. Pipelines	17
3.2.6. Mapping Data Flow	18
3.3. Biến đổi dữ liệu với Azure Databricks	18
3.3.1. Giới thiệu	18
3.3.2. Thành phần	19
3.3.3. Đặc điểm	20
3.4. Azure SQL Database	20
3.4.1. Giới thiệu	21
3.4.2. Mô hình triển khai	21
3.4.3. Đặc điểm	21
3.4.4. Sử dụng Azure SQL Database làm Data warehouse.....	22
3.5. Trực quan hóa dữ liệu bằng Power BI service	23
3.5.1. Giới thiệu	23
3.5.2. Direct Query.....	23
3.5.3. Row-level security	24
3.5.4. Đặc điểm khác	24
III. MÔ HÌNH DỮ LIỆU - TRIỂN KHAI	25

1. Kết quả triển khai mô hình.....	25
1.1. Luồng xử lý dữ liệu tự động ETL	25
1.1.1. Data pipeline.....	25
1.1.2. Định dạng lưu trữ	28
1.1.3. Đo tốc độ thực thi.....	28
1.1.4. Đo độ trễ khi thiết lập server ở 2 vùng khác nhau	29
1.2. Hiện thực trực quan hóa dữ liệu.....	29
1.2.1. Tổng quan dashboard	29
1.2.2. Các chức năng chi tiết đã hiện thực trên dashboard.....	32
1.3. Giải pháp tối ưu luồng xử lý dữ liệu.....	36
2. Chi tiết hiện thực triển khai	36
2.1. Luồng xử lý dữ liệu tự động ETL	36
2.1.1. Cài đặt Resource Group.....	36
2.1.2. Cài đặt Storage Account	38
2.1.3. Cài đặt SQL Database and SQL server.....	40
2.1.4. Cài đặt Azure Databricks	45
2.1.5. Cài đặt Azure Data Factory.....	49
2.2. Hiện thực trực quan hóa dữ liệu	55
2.3. Giải pháp tối ưu	66
THAM KHẢO.....	70

I. GIỚI THIỆU BÀI TOÁN

1. Tổng quan bài toán

Lending Club, một trong những nền tảng cho vay ngang hàng lớn nhất tại Hoa Kỳ, đã nổi tiếng với mô hình kinh doanh P2P (Peer-to-Peer), cho phép người vay vốn kết nối trực tiếp với các nhà đầu tư cá nhân hoặc tổ chức. Kể từ khi thành lập vào năm 2007, Lending Club đã tạo ra một hệ thống trực tuyến linh hoạt và tiện lợi, đóng vai trò là cầu nối giữa người vay và người cho vay.



Hình 1. Tổng quan về Công ty Lending Club

- Cách hoạt động của Lending Club

- + **Đăng ký và Xác thực:** Người vay đăng ký tài khoản trên nền tảng của Lending Club và cung cấp thông tin cá nhân và tài chính của mình. Lending Club tiến hành quá trình xác thực và đánh giá rủi ro tín dụng của người vay để đảm bảo tính minh bạch và an toàn cho cả bên vay và bên cho vay.
- + **Đăng Ký Vay:** Người vay chọn số tiền muốn vay và mục đích sử dụng vốn. Họ cung cấp thông tin về thu nhập, nợ nần và các thông tin tài chính khác để giúp Lending Club hiểu rõ hơn về tình hình tài chính của họ.

- + **Xác Nhận Tín Dụng:** Dựa trên dữ liệu tài chính được cung cấp, Lending Club xác định khả năng trả nợ của người vay và xác định mức lãi suất phù hợp cho khoản vay.
 - + **Cung cấp cho nhà đầu tư:** Sau khi được xác nhận, các khoản vay được đưa ra cho nhà đầu tư trên nền tảng của Lending Club. Nhà đầu tư có thể đầu tư vào một phần hoặc toàn bộ của khoản vay, tạo ra một cơ hội đầu tư đa dạng và tiềm năng lợi nhuận cao.
 - + **Quản lý và Trả nợ:** Người vay thực hiện việc trả nợ thông qua các khoản trả hàng tháng, bao gồm cả lãi suất và gốc. Lending Club đảm bảo quản lý hiệu quả quá trình thu nợ và chia lợi tức cho nhà đầu tư một cách công bằng và minh bạch.
 - + **Quản lý Rủi ro:** Lending Club thực hiện các biện pháp quản lý rủi ro như theo dõi định kỳ các tình hình tài chính, xử lý các khoản nợ xấu và tuân thủ nghiêm ngặt các quy định về tài chính và cho vay.
- **Yêu cầu từ phía công ty,** Công ty Lending Club đặt ra các yêu cầu nhất định để quản lý và theo dõi hoạt động cho vay của mình:



Hình 2. *Đưa ra quyết định dựa trên dữ liệu của doanh nghiệp*

- + **Theo dõi tình hình cho vay theo thời gian:** Công ty cần theo dõi số lượng và giá trị của các khoản vay được cấp trong các khoảng thời gian khác nhau (hàng

tháng, hàng quý, hàng năm) để đảm bảo sự ổn định và tính minh bạch của hoạt động kinh doanh.

- + **Đánh giá hiệu quả hoạt động của các khoản vay:** Công ty cần đánh giá hiệu quả của từng khoản vay dựa trên các tiêu chí như số tiền vay, lãi suất, thời hạn vay và xếp hạng tín dụng của người vay. Điều này giúp họ hiểu rõ hơn về lợi ích và rủi ro liên quan đến từng khoản vay.
- + **Phân tích xu hướng và mẫu hình trong dữ liệu:** Công ty cần phân tích các xu hướng và mẫu hình trong dữ liệu để dự đoán các mô hình cho vay trong tương lai. Việc này có thể bao gồm việc xác định xu hướng tăng hoặc giảm của các khoản vay theo mùa hoặc nhận định các yếu tố ảnh hưởng đến tỷ lệ hoàn trả của khoản vay.

⇒ **Mục tiêu bài toán đề ra:** Mục tiêu của bài toán là xử lý, lưu trữ và trực quan hóa dữ liệu để cung cấp thông tin hữu ích giúp công ty quản lý và ra quyết định hiệu quả hơn về hoạt động cho vay. Điều này giúp Lending Club cải thiện hiệu suất và giảm thiểu rủi ro trong việc cho vay, tạo ra một môi trường kinh doanh ổn định và bền vững.

2. Tổng quan về bộ dữ liệu

- **Loại dữ liệu:** Dữ liệu về các khoản vay của nền tảng cho vay ngang hàng (P2P – Lending club) lớn nhất từ năm 2007 đến quý 3 năm 2020 với quy mô 2,9 triệu dòng và 141 cột.
- **Kích thước dữ liệu:** 1,73 GB
- **Kiểu dữ liệu:** CSV - Comma-Separated Values
- **Nguồn thu thập:** Kaggle -

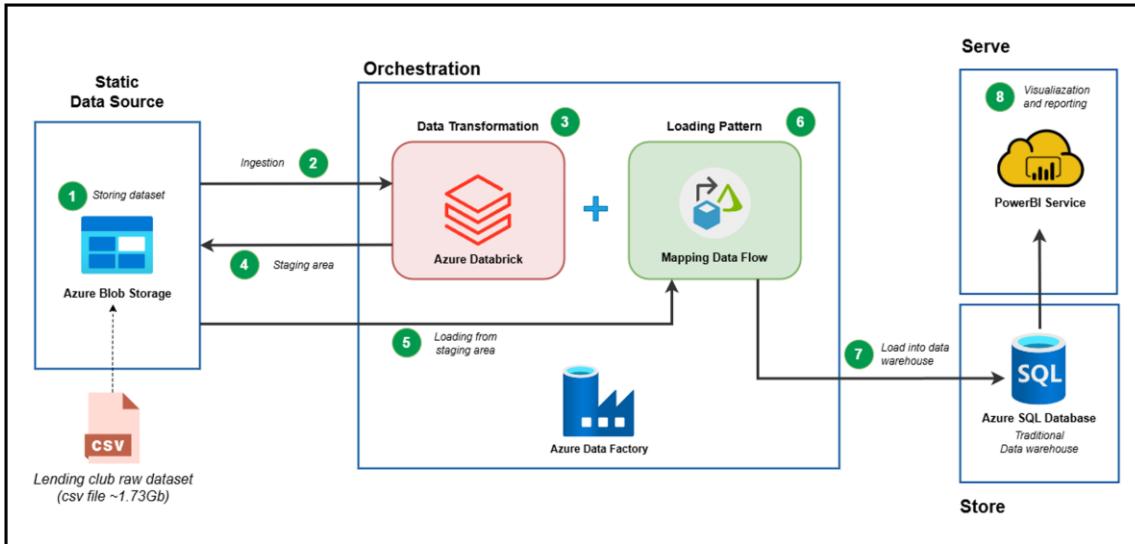
<https://www.kaggle.com/datasets/ethon0426/lending-club-20072020q1>

3. So sánh tốc độ xử lý máy chủ truyền thống và hệ thống cloud

Đặc điểm	Truyền thống	Cloud
Tốc độ phản hồi dữ liệu	Thường lưu trữ và xử lý các dữ liệu cục bộ, ít gây ra sự chậm trễ	Dữ liệu cần phải truy cập qua mạng trước khi được trả về, gây ra sự chậm trễ
Chia sẻ dữ liệu	Không chia sẻ tài nguyên được với các máy chủ khác	Cho phép kết nối, phân tán dữ liệu cho nhiều máy chủ
Đổi mới và linh hoạt	Cần thời gian để nâng cấp và tích hợp	Nhanh chóng áp dụng công nghệ mới, triển khai và thử nghiệm ứng dụng nhanh, dễ dàng tích hợp công cụ hơn
Khả năng mở rộng quy mô	Tốn nhiều thời gian để nâng cấp	Dễ dàng mở rộng quy mô quản lý dữ liệu
Độ ổn định	Máy chủ hỏng sẽ dẫn đến toàn bộ hệ thống ngừng hoạt động.	Các thành phần được thiết lập dự phòng đảm bảo khôi phục sự cố, dữ liệu gần như được truy cập mọi lúc mọi nơi

Bảng 1. So sánh tốc độ xử lý giữa máy chủ truyền thống và hệ thống cloud

4. Kiến trúc dự án



Hình 3. Kiến trúc tổng quan của hệ thống

4.1. Phân loại các thành phần trong kiến trúc

- PaaS - Platform as a Service:

- + Azure Blob Storage
- + Azure Data Factory
- + Azure Databricks
- + Mapping Data Flow
- + Azure SQL Database

- SaaS - Software as a Service:

- + PowerBI Service

4.2. Chức năng của các thành phần trong kiến trúc

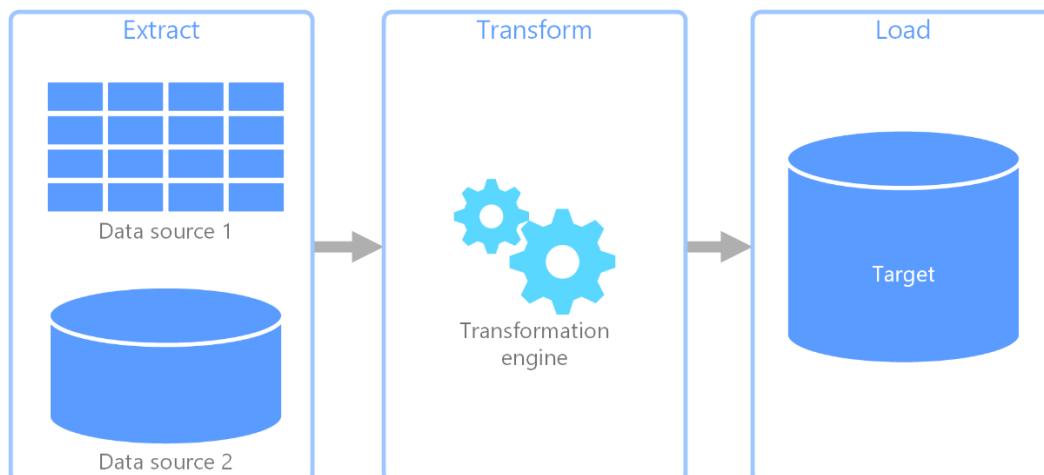
- Azure Blob Storage: Sử dụng để lưu trữ dữ liệu thô chưa xử lý, dữ liệu được thu thập từ kaggle và upload lên dịch vụ này.
- Azure Data Factory: Công cụ chính cho việc điều phối và tự động hóa luồng dữ liệu, hỗ trợ thiết kế luồng xử lý dữ liệu tự động trên môi trường cloud. Hỗ trợ đa dạng loại kết nối đến các dịch vụ trong kiến trúc đề ra.

- Azure Databricks: Nền tảng phân tích dữ liệu phân tán, thực hiện xử lý, làm sạch và cấu trúc dữ liệu thô ở quy mô lớn thông qua ngôn ngữ lập trình python. Sử dụng làm 1 thành phần tích hợp trong data pipeline của Azure Data Factory.
- Mapping Data Flow: Công cụ hỗ trợ thực hiện tổ chức và mô hình hóa dữ liệu dưới dạng star schema để đưa dữ liệu vào Data warehouse.
- Azure SQL Database: Sử dụng Azure SQL Database (SQL Server) làm Data warehouse truyền thống để tổ chức và lưu trữ dữ liệu phục vụ cho việc truy vấn và phân tích có cấu trúc. Azure SQL Database phù hợp hơn với các tập dữ liệu quan hệ và dung lượng dữ liệu nhỏ (<4TB).
- PowerBI Service: Công cụ phân tích dữ liệu kinh doanh theo dạng dịch vụ (SaaS), giúp tạo báo cáo và hình ảnh trực quan từ dữ liệu đã qua xử lý trong Azure SQL Database. Chia sẻ và quản lý các bảng báo cáo rộng rãi đến đa dạng thiết bị và người dùng có kết nối đến với Internet.

II. CƠ SỞ LÝ THUYẾT

1. Định dạng lưu trữ

1.1. Quá trình ETL (Extract, Transform, Load)



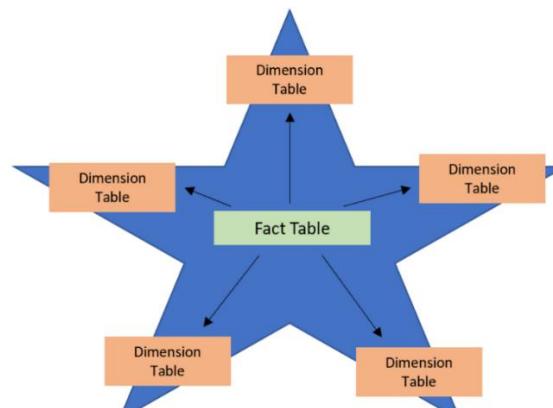
Hình 4. Quá trình ETL.

Nguồn: [Extract, transform, load \(ETL\) - Azure Architecture Center / Microsoft Learn](#)

Quá trình ETL (Extract, Transform, Load) là một quy trình quan trọng trong việc xây dựng các hệ thống kho dữ liệu (Data Warehouse). Nó bao gồm ba bước chính:

- **Extract (Trích xuất):** Đây là giai đoạn đầu tiên của quá trình ETL, trong đó dữ liệu được trích xuất từ các nguồn dữ liệu khác nhau như cơ sở dữ liệu quan hệ, hệ thống ERP, tệp tin, dịch vụ web, và các nguồn dữ liệu khác. Mục tiêu của giai đoạn này là thu thập dữ liệu cần thiết từ các hệ thống khác nhau mà không làm ảnh hưởng đến hoạt động của chúng.
- **Transform (Biến đổi):** Sau khi dữ liệu được trích xuất, nó cần được biến đổi để phù hợp với mô hình dữ liệu của kho dữ liệu. Quá trình biến đổi có thể bao gồm các bước như:
 - + **Làm sạch dữ liệu (Data Cleaning):** Loại bỏ hoặc sửa các dữ liệu bị lỗi, không nhất quán hoặc thiếu.
 - + **Chuyển đổi định dạng (Data Transformation):** Thay đổi định dạng của dữ liệu để phù hợp với yêu cầu của kho dữ liệu. Ví dụ, chuyển đổi định dạng ngày tháng, hợp nhất các trường dữ liệu, v.v.
 - + **Tích hợp dữ liệu (Data Integration):** Kết hợp dữ liệu từ nhiều nguồn khác nhau vào một cấu trúc nhất quán.
- **Load (Tải):** Giai đoạn cuối cùng là tải dữ liệu đã biến đổi vào kho dữ liệu. Quá trình này phải đảm bảo rằng dữ liệu được tải vào đúng cách, không gây xung đột và duy trì tính toàn vẹn của dữ liệu.

1.2. Mô hình hóa dữ liệu – Star schema



tomanhhoang.com

Hình 5. Mô hình Star schema.

Nguồn: [Mô hình hóa dữ liệu: Star Schema, Bảng DIM và bảng FACT \(tomanhhoang.com\)](http://tomanhhoang.com)

Mô hình Star Schema (ngôi sao) là một trong những kiến trúc phổ biến nhất được sử dụng trong các hệ thống kho dữ liệu. Mô hình này được gọi là "ngôi sao" vì cấu trúc của nó có một bảng sự kiện (fact table) ở trung tâm và các bảng chiều (dimension tables) bao quanh, tạo thành hình dạng giống như một ngôi sao.

- Thành phần của Star Schema:
 - + **Fact Table (Bảng sự kiện):** Đây là bảng trung tâm chứa các dữ liệu đo lường hoặc sự kiện thực tế, thường bao gồm các số liệu cần phân tích như số lượng, doanh thu, chi phí, v.v. Mỗi bản ghi trong bảng sự kiện thường có khóa ngoại liên kết đến các bảng chiều.
 - + **Dimension Tables (Bảng chiều):** Các bảng này chứa dữ liệu mô tả hoặc phân loại các khía cạnh khác nhau của dữ liệu trong bảng sự kiện. Ví dụ, trong một hệ thống theo dõi bán hàng, các bảng chiều có thể bao gồm "Thời gian", "Sản phẩm", "Khách hàng", và "Địa điểm". Mỗi bảng chiều thường có một khóa chính (primary key) được tham chiếu bởi khóa ngoại trong bảng sự kiện
- Lợi ích của Star Schema:
 - + **Hiệu suất truy vấn cao:** Mô hình Star Schema được tối ưu hóa cho các truy vấn OLAP (Online Analytical Processing), cho phép truy xuất dữ liệu nhanh chóng và hiệu quả.
 - + **Dễ hiểu và dễ quản lý:** Cấu trúc đơn giản và trực quan giúp dễ dàng quản lý và mở rộng.
 - + **Hỗ trợ phân tích đa chiều:** Cho phép phân tích dữ liệu từ nhiều góc độ khác nhau, phù hợp cho các báo cáo và phân tích phức tạp.

1.3. Thiết kế kho dữ liệu với sự kết hợp ETL và mô hình Star schema

- **Phân tích yêu cầu và thu thập dữ liệu**
 - + Xác định các chỉ số kinh doanh: Bước đầu tiên là xác định các chỉ số kinh doanh quan trọng (KPIs) mà công ty muốn theo dõi, chẳng hạn như doanh

thu, chi phí, số lượng hợp đồng cho thuê, tần suất khách hàng quay lại, v.v.

- + Các nguồn dữ liệu cần tích hợp: Liệt kê các nguồn dữ liệu hiện có, bao gồm cơ sở dữ liệu giao dịch, hệ thống ERP, tệp tin Excel, dịch vụ web và các nguồn khác.

- **Thiết kế mô hình Star Schema**

- + Xác định bảng sự kiện (Fact Table): Xác định các sự kiện hoặc giao dịch chính cần theo dõi trong hệ thống.
- + Xác định bảng chiều (Dimension Tables): Thiết kế các bảng chiều để chứa thông tin mô tả về các khía cạnh của bảng sự kiện.

- **Thiết kế và triển khai quá trình ETL**

Trích xuất dữ liệu từ các nguồn

- + Thu thập dữ liệu: Trích xuất dữ liệu từ các hệ thống nguồn khác nhau như cơ sở dữ liệu giao dịch, hệ thống ERP, tệp tin Excel, dịch vụ web, đảm bảo không ảnh hưởng đến hoạt động của hệ thống nguồn.

Biến đổi dữ liệu để làm sạch và chuẩn hóa

- + Làm sạch dữ liệu: Xử lý các dữ liệu bị thiếu, không hợp lệ hoặc không nhất quán. Ví dụ, loại bỏ các bản ghi trùng lặp hoặc sửa lỗi chính tả trong các trường dữ liệu.
- + Chuẩn hóa dữ liệu: Chuyển đổi dữ liệu về định dạng thống nhất, chẳng hạn như chuẩn hóa định dạng ngày tháng, hợp nhất các trường dữ liệu.
- + Tích hợp dữ liệu: Kết hợp dữ liệu từ nhiều nguồn khác nhau để tạo thành một bộ dữ liệu nhất quán và chuẩn hóa theo mô hình Star Schema.

Tải dữ liệu vào các bảng sự kiện và bảng chiều trong kho dữ liệu

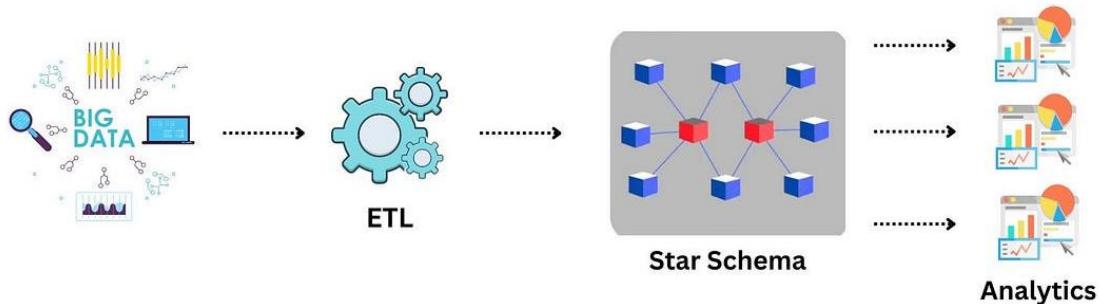
- + Tải dữ liệu: Đưa dữ liệu đã biến đổi vào các bảng sự kiện và bảng chiều trong kho dữ liệu. Đảm bảo tính toàn vẹn và nhất quán của dữ liệu trong quá trình tải.

- **Triển khai và tối ưu hóa**

- + Công cụ BI (Business Intelligence): Sử dụng các công cụ BI như Power BI, Tableau, hoặc các công cụ phân tích dữ liệu khác để kết nối với kho dữ liệu, tạo các báo cáo và bảng điều khiển (dashboards).
- + Tối ưu hóa hiệu suất: Áp dụng các kỹ thuật tối ưu hóa như indexing, partitioning, và caching để cải thiện hiệu suất truy vấn và đảm bảo hệ thống hoạt động mượt mà.

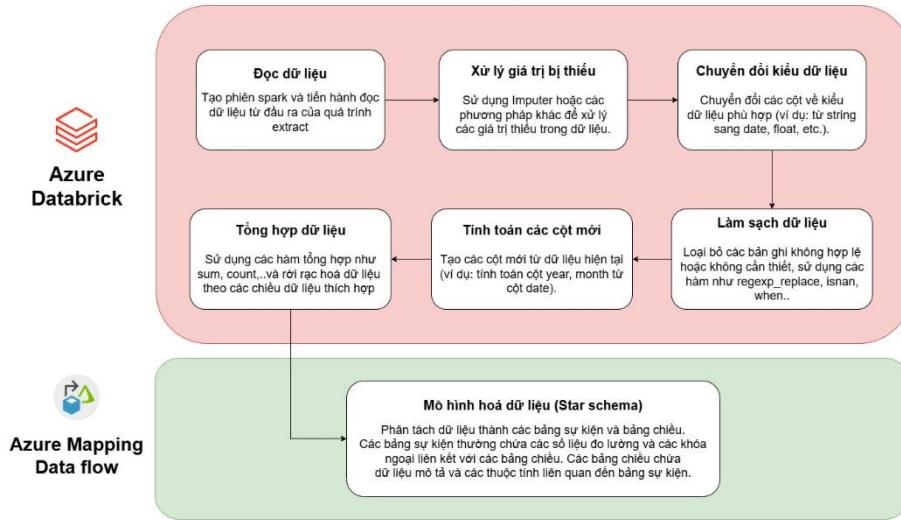
- **Quản lý và bảo trì**

- + Cập nhật quy trình ETL: Thường xuyên cập nhật quy trình ETL để phản ánh các thay đổi trong nguồn dữ liệu hoặc yêu cầu kinh doanh mới.
- + Theo dõi hiệu suất: Liên tục theo dõi hiệu suất của kho dữ liệu và quy trình ETL, phát hiện và khắc phục các vấn đề hiệu suất kịp thời.
- + Bảo mật dữ liệu: Đảm bảo dữ liệu trong kho được bảo mật, tuân thủ các quy định về bảo vệ dữ liệu và quyền riêng tư, ví dụ như sử dụng các cơ chế mã hóa và kiểm soát truy cập chặt chẽ.



Hình 6. Sự kết hợp giữa quá trình ETL và mô hình Star schema.
Nguồn: [Data Modelling: Techniques, Importance and Implementation / by Venkatakrishnan / Medium](#)

2. Kỹ thuật xử lý dữ liệu



Hình 7. Tổng quan kỹ thuật xử lý dữ liệu được sử dụng

2.1. Đọc dữ liệu

- Tạo notebook và thiết lập phiên spark, tiến hành đọc dữ liệu từ đầu ra của quá trình extract trong Azure Data Factory.
- Đọc dữ liệu và sê thao tác dưới định dạng schema – Spark dataframe

2.2. Xử lý giá trị bị thiếu

- Loại bỏ các bản ghi chứa giá trị bị thiếu trầm trọng và không ảnh hưởng đến kết quả cuối cùng
- Sử dụng Imputer, điền các dữ liệu bị thiếu bằng các giá trị trung vị hoặc trung bình của cột đó

2.3. Chuyển đổi kiểu dữ liệu

- Chuyển đổi kiểu dữ liệu về định dạng và kiểu dữ liệu phù hợp
- Các kiểu dữ liệu cần nhắc chuyển đổi như: từ string sang date/datetime, float, numeric, string và kiểu dữ liệu phần trăm..

2.4. Làm sạch dữ liệu

- Loại bỏ các bản ghi không hợp lệ hoặc không cần thiết.
- Sử dụng các hàm như regexp_replace, isnan, when.. để thao tác làm sạch dữ liệu.

2.5. Tính toán các cột mới

- Tạo các cột dữ liệu mới từ các thông tin hiện có trong bộ dữ liệu.
- Tính toán IRR, ROI, FICO của từng khoản vay nhằm mục đích lưu trữ và phân tích về sau.
- Tính toán các giá trị dti, int_rate.. để kiểm tra các dữ liệu đang có sẵn trong bộ dữ liệu.

2.6. Tổng hợp dữ liệu

- Sử dụng các hàm tổng hợp dữ liệu như sum, count, min/max,..
- Rời rạc hóa dữ liệu đối với các dữ liệu dạng dữ liệu số liên tục.

2.7. Mô hình hoá dữ liệu (Star schema)

- Phân loại dữ liệu thành các bảng sự kiện và bảng chiều
- Các bảng sự kiện chứa số liệu đo lường và các khoá ngoại liên kết đến các bảng chiều.
- Các bảng chiều biểu diễn các đặc trưng bao gồm dữ liệu mô tả và thuộc tính liên quan đến bảng sự kiện.

2.8. Kết hợp sử dụng Azure Databricks và Mapping data flow cho việc biến đổi dữ liệu

- Sử dụng cả Azure Databricks và Mapping Data Flow trong Azure Data Factory là một chiến lược tổng thể mạnh mẽ để xử lý dữ liệu. Kết hợp sức mạnh tính toán và linh hoạt của Azure Databricks với tính dễ sử dụng và tự động hóa của Mapping Data Flow cho phép tối ưu hóa hiệu suất, giảm thiểu chi phí, và đáp ứng mọi nhu cầu từ các tác vụ đơn giản đến những nhiệm vụ phức tạp nhất trong quy trình xử lý dữ liệu.
- Phân loại các công việc biến đổi phức tạp và khôi lượng lớn để tận dụng tối đa sức mạnh của ngôn ngữ lập trình để biến đổi dữ liệu bằng Azure Databricks
- Phân loại các công việc biến đổi tổ chức mô hình hóa dữ liệu (Star schema) sẽ sử dụng Mapping Data Flow, để có được sự kiểm soát về toàn bộ tổ chức dữ liệu và kiểu dữ liệu thông qua điểm mạnh về biến đổi dữ liệu thông qua giao diện trực quan của công cụ này

3. Các thành phần

3.1. Sử dụng Azure Blob Storage để lưu trữ dữ liệu nguồn

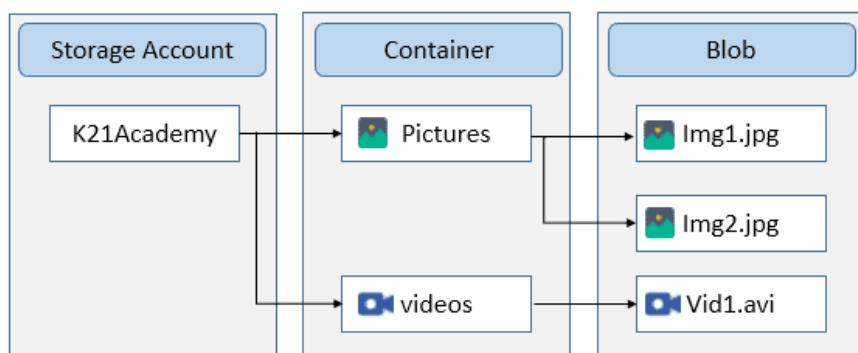
3.1.1. Giới thiệu

Blob (Binary Large Object) Storage là dịch vụ lưu trữ đối tượng trên đám mây của Microsoft Azure. Nó cho phép các nhà phát triển lưu trữ lượng lớn dữ liệu phi cấu trúc trong nền tảng đám mây của Microsoft và truy cập tới các dữ liệu này từ mọi lúc, mọi nơi thông qua HTTP hoặc HTTPS..

Tốc độ đọc ghi: Phụ thuộc vào loại blob, kích thước, loại storage account, số lượng request,...

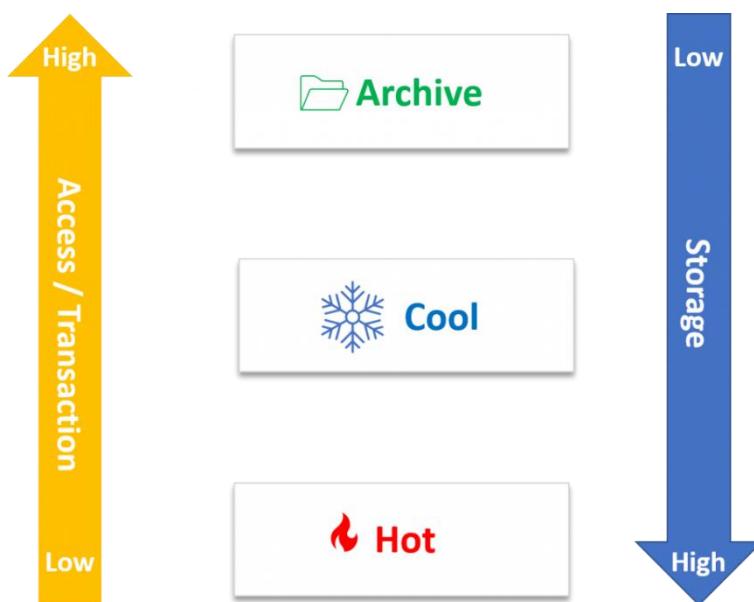
3.1.2. Các thành phần liên quan

- Azure Blob Storage cung cấp 3 loại tài nguyên:
 - + **Storage account**: Đây là nơi bạn tạo và quản lý tất cả các tài nguyên liên quan đến lưu trữ blob. Mỗi tài khoản lưu trữ có một tên duy nhất và cung cấp một không gian tên riêng để lưu trữ dữ liệu.
 - + **Container**: Đây là một phần của tài khoản lưu trữ và hoạt động tương tự như một thư mục trong hệ thống tệp.
 - + **Blob**: Đây là đơn vị cơ bản của dữ liệu lưu trữ trong Blob Storage. Các blob được lưu trữ trong các container và được truy cập thông qua các URL duy nhất.



Hình 8. Thành phần của Blob Storage. Nguồn: k21academy.com

- Chi phí lưu trữ dữ liệu phụ thuộc vào hai yếu tố: chi phí giao dịch và chi phí lưu trữ. Azure Blob Storage cung cấp các tầng truy cập khác nhau để lưu trữ dữ liệu blob của bạn tùy thuộc vào cách sử dụng của nó
- + **Tầng nóng (Hot tier):** Chi phí lưu trữ cao nhưng chi phí giao dịch thấp. Phù hợp để lưu trữ dữ liệu mà thường xuyên được truy cập hoặc chỉnh sửa.
- + **Tầng mát (Cool tier):** Cả hai chi phí lưu trữ thấp nhưng chi phí truy cập cao hơn tầng nóng. Phù hợp để lưu trữ những dữ liệu ít được truy cập không cần có sẵn hoặc sao lưu ngắn hạn.
- + **Tầng lưu trữ (Archive):** Chi phí lưu trữ rất thấp nhưng chi phí truy cập cao. Sử dụng để lưu trữ các dữ liệu hiếm được truy cập nhưng cần giữ để lưu trữ hoặc sao lưu dài hạn.



Hình 9. Ba tầng truy cập của Blob Storage. Nguồn: linkedin.com

3.1.3. Đặc điểm

- Lưu trữ theo tầng, chi phí thấp
- Tính sẵn sàng cao
- Tính nhất quán mạnh mẽ
- Khả năng khắc phục thảm họa

3.2. Điều phối bằng Azure Data Factory

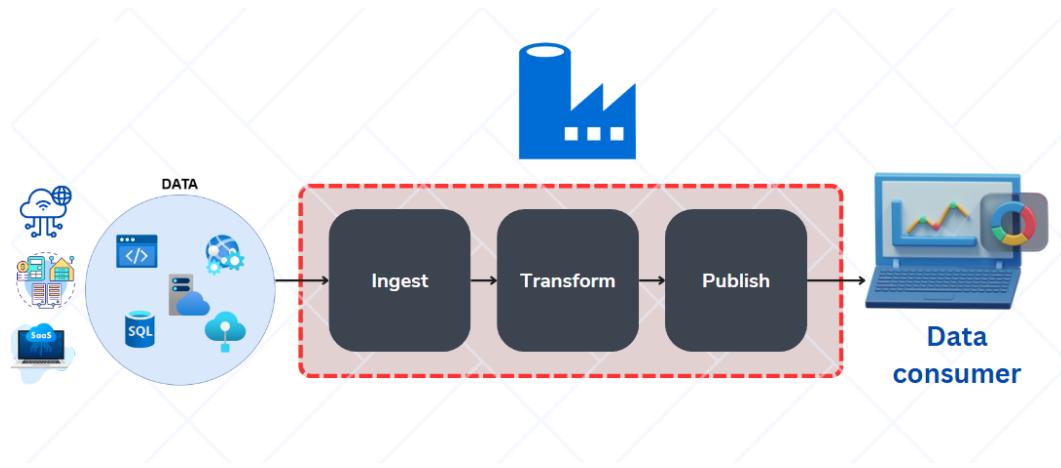
Azure Data Factory là dịch vụ tích hợp dữ liệu cloud-based cho phép tạo quy trình làm việc dựa trên dữ liệu trên đám mây để điều phối và tự động hóa việc di chuyển và chuyển đổi dữ liệu.

- + **Thu thập (Ingest):** Azure Data Factory hỗ trợ thu thập dữ liệu từ nhiều nguồn
- + **Biến đổi (Transform):** Azure Data Factory cung cấp khả năng biến đổi dữ liệu thông qua

Dataflow: cung cấp các tính năng transform dữ liệu, tuy nhiên vẫn có một số hạn chế về mặt xử lý các tác vụ biến đổi dữ liệu phức tạp.

Thay vào đó, Azure Data Factory cho phép kết hợp biến đổi dữ liệu với các công cụ từ dịch vụ khác như Databricks, HDInsight..

- + **Xuất bản (Publish):** Hỗ trợ để đưa các dữ liệu sau khi xử lý vào các công cụ khác như PowerBI, Azure SQL Database, Azure Synapse,... để phân tích và sử dụng dữ liệu.



Hình 10. Azure Data Factory - Usecase

Azure Data Factory là dịch vụ ETL (Extract, Transform, Load) trên nền tảng đám mây của Microsoft Azure và tích hợp dữ liệu cho phép tạo ra các quy trình tự động để điều phối việc di chuyển dữ liệu và chuyển đổi dữ liệu trên quy mô lớn.

Tốc độ đọc ghi: Phụ thuộc vào băng thông mạng, số lượng copy activity, cấu hình nguồn/đích,...

3.2.1. Extract

- Quá trình trích xuất này bao gồm việc thu thập dữ liệu từ nhiều nguồn như cơ sở dữ liệu, lưu trữ đám mây,...cần xác định dữ liệu (Data) và nguồn của nó (Data Source).
 - + Data source: Xác định các chi tiết nguồn như đăng ký, nhóm tài nguyên và thông tin nhận dạng như khóa.
 - + Data: Xác định dữ liệu bằng cách sử dụng một bộ tệp, truy vấn cơ sở dữ liệu hoặc tên lưu trữ Azure Blob cho lưu trữ blob.

3.2.2. Transform

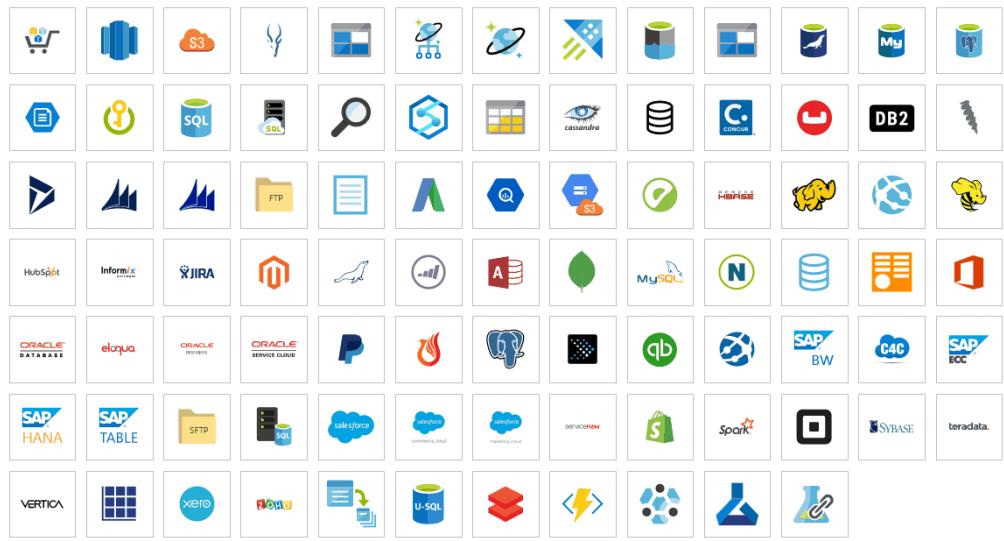
- Dữ liệu sau khi được trích xuất thường yêu cầu làm sạch, lọc, định dạng lại và kết hợp với các bộ dữ liệu khác để đảm bảo tính nhất quán và chất lượng. Các hoạt động chuyển đổi dữ liệu có thể bao gồm kết hợp, chia tách, thêm, lấy, xóa hoặc xoay cột.
- Ánh xạ các field giữa dữ liệu đích (Data Destination) và dữ liệu nguồn.

3.2.3. Load

- Sau khi dữ liệu được chuyển đổi và chuẩn bị, nó sẽ được tải vào hệ thống mục tiêu hoặc đích; có thể là kho dữ liệu (Data warehouse), cơ sở dữ liệu (Database), hồ dữ liệu (Data Lake) hoặc bất kỳ cơ sở hạ tầng lưu trữ nào khác được tối ưu hóa cho phân tích hoặc báo cáo.

3.2.4. Linked Services

Linked service (Các dịch vụ liên kết) thiết lập kết nối với các kho dữ liệu bên ngoài. Chúng bao gồm thông tin kết nối và thông tin xác thực, cho phép Azure Data Factory truy cập và truy xuất dữ liệu từ các nguồn khác nhau một cách an toàn.

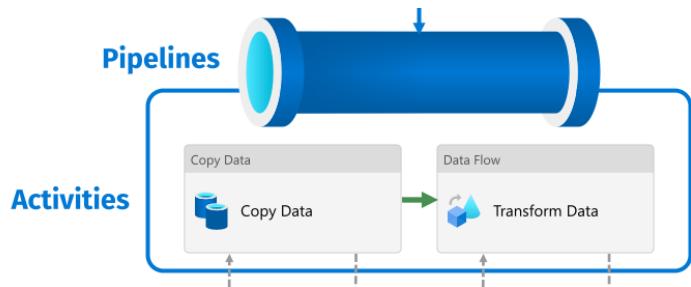


© 2019 Cathrine Wilhelmsen (hi@cathrinewilhelmsen.net)

Hình 11. Các Linked Services ví dụ. Nguồn: cathrinewilhelmsen.net

3.2.5. Pipelines

- Một Azure Data Factory có thể có một hoặc nhiều pipelines.
- Pipeline đại diện cho một chuỗi các hoạt động có tính liên kết với nhau, hoặc có thể hiểu đó là một quy trình xuyên suốt và vận hành liên tục.
- Một Data Pipeline sẽ thực hiện các nhiệm vụ tổng hợp, sắp xếp và di chuyển dữ liệu đến hệ thống mục tiêu nhằm tiến hành lưu trữ và phân tích.
- Activity (Hoạt động) đại diện bước xử lý nhỏ trong pipeline. Azure Data Factory cung cấp các loại Activity như:
 - + Di chuyển dữ liệu (Copy activity)
 - + Biến đổi dữ liệu (Data flow activity)
 - + Kiểm soát dòng chảy dữ liệu (Data control activity)



Hình 12. Minh họa giữa Activities và Pipelines. Nguồn: Cagthrine Wilhelmsen (hi@cathrinew.net)

3.2.6. Mapping Data Flow



Hình 13. Mapping Data Flows. Nguồn: clearpeaks.com

Trong Azure Data Factory, một Data flow là một activity có thể được thêm vào Pipeline. Data Flow activity được sử dụng để chuyển dữ liệu từ nguồn đến đích sau khi thực hiện một số chuyển đổi trên dữ liệu như join, aggregate, filter, và pivot.

Mapping data flows là các phép biến đổi dữ liệu được thiết kế trực quan, giúp phát triển logic chuyển đổi dữ liệu mà không cần sử dụng đến code. Mapping data flow sau khi được tạo và kiểm tra, có thể được thêm vào Data Flow activity của Pipeline. Data Factory xử lý tất cả việc dịch code, tối ưu hóa đường dẫn và thực thi các Data Flow.

Tốc độ đọc ghi: Phụ thuộc vào kích thước, nguồn đọc dữ liệu, độ phức tạp của data flow,...

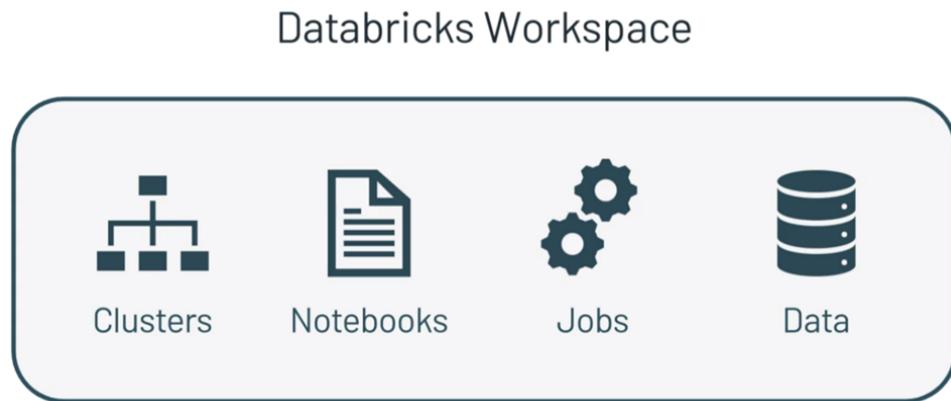
3.3. Biến đổi dữ liệu với Azure Databricks

3.3.1. Giới thiệu

Azure Databricks là một nền tảng phân tích dữ liệu, dựa trên Apache Spark, được xây dựng trên nền tảng đám mây Microsoft Azure, cho phép người dùng dễ dàng tạo một không gian làm việc, triển khai, chia sẻ và duy trì dữ liệu.

Tốc độ đọc ghi: Phụ thuộc vào cấu hình cluster, loại lưu trữ,...

3.3.2. Thành phần



Hình 14. Thành phần cơ bản trong Databricks. Nguồn: [faun.pub](#)

- **Workspace:** là một nơi có thể quản lý tất cả dữ liệu hoặc tệp tin theo định dạng thư mục, có thể là các Notebook, các thư viện khác nhau, bảng điều khiển trực quan, thử nghiệm ML,... Người quản lý có thể thiết lập quyền truy cập cho từng thành viên khác trong Workspace.
- **Cluster:** Đây là thành phần quan trọng nhất trong Database và Spark để thực thi dữ liệu ở một không gian nhanh hơn rất nhiều. Có hai loại cluster có thể tạo trong Databricks:
 - + Cluster Tương tác (Interactive/All-Purpose Cluster): cho phép nhiều người dùng khám phá và phân tích dữ liệu theo cách tương tác
 - + Cluster Công việc (Job Cluster): được sử dụng để chạy các công việc nhanh chóng và tự động.
- **Notebook:** là một phiên bản của Jupyter notebook, là công cụ cho phép bạn viết và thực thi các đoạn code. Hoặc nó có thể thực hiện các biến đổi dữ liệu khác nhau trên dữ liệu với các ngôn ngữ được hỗ trợ bởi Spark. Trong Notebook của DataBricks, code có thể được viết bằng Python, SQL, Scala, R và trong cùng một Notebook có thể code bằng các ngôn ngữ khác nhau này.
- **Workspace:** là một nơi có thể quản lý tất cả dữ liệu hoặc tệp tin theo định dạng thư mục, có thể là các Notebook, các thư viện khác nhau, bảng điều khiển trực

quan, thử nghiệm ML,... Databricks hỗ trợ xác định kiểm soát truy cập chi tiết trên tất cả các đối tượng này, cho phép người dùng sử dụng linh hoạt cùng một Workspace, nhưng chỉ cung cấp hạn chế quyền truy cập cho họ.

- **Job:** Job cho phép thực thi một Notebook. Một Job có thể chạy ngay lập tức hoặc có thể được lập lịch. Và Job có thể chạy trên các Job cluster.

- Ví dụ, nếu một tệp JAR bên ngoài mà muốn thực thi trên một Cluster Spark. Ta có thể làm điều đó bằng cách sử dụng các Job.

3.3.3. Đặc điểm

- **Xử lý Big Data:** Azure Databricks hỗ trợ xử lý dữ liệu lớn bằng cách sử dụng Apache Spark, cho phép người dùng thực hiện các tác vụ như xử lý dữ liệu cấu trúc và không cấu trúc, thống kê, và phân tích dữ liệu với hiệu suất cao.
- **Phân tích thời gian thực:** Có thể được sử dụng để phân tích dữ liệu luồng trong thời gian thực, cho phép các tổ chức thu thập thông tin và hành động nhanh chóng.
- **Môi trường cộng tác:** Cung cấp một môi trường cộng tác cho phép các nhóm làm việc cùng nhau và chia sẻ notebook, dữ liệu, kiến thức qua các dự án.
- **Tự động hóa:** Nó cung cấp các tính năng tự động hóa giúp đơn giản hóa việc tạo, quản lý và triển khai các khối công việc xử lý dữ liệu lớn và máy học. Nó cung cấp các tính năng cung cấp cụm tự động, tự động mở rộng và lập lịch công việc.
- **Tích hợp:** Azure Databricks tích hợp với nhiều dịch vụ Azure khác như Azure Data Factory, Azure Event Hubs và Azure Blob Storage. Điều này cho phép các nhóm dễ dàng xây dựng các đường ống dữ liệu đầu cuối để thu nạp, xử lý và phân tích dữ liệu trong thời gian thực.
- **Bảo mật:** Nó cung cấp các tính năng bảo mật mạnh mẽ như kiểm soát truy cập dựa trên vai trò (role-based access control), cô lập mạng (network isolation) và mã hóa dữ liệu (data encryption). Điều này đảm bảo cho dữ liệu của các tổ chức được an toàn và bảo mật.

3.4. Azure SQL Database

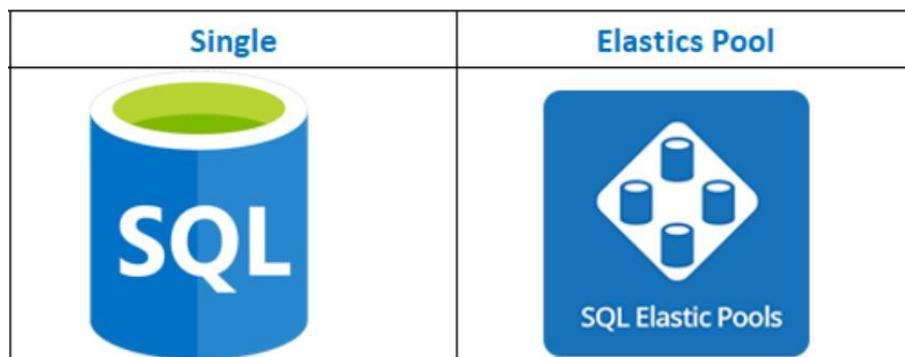
3.4.1. Giới thiệu

Azure SQL Database là một cơ sở dữ liệu đám mây được quản lý, dựa trên Microsoft SQL Server, và là một phần của các dịch vụ Microsoft Azure, cung cấp các chức năng quản lý cơ sở dữ liệu tự động và cần rất ít sự tham gia của người dùng như update, patch, backups, và monitor.

Khi được triển khai làm Data Warehouse, Azure SQL Database cho phép tổ chức dữ liệu theo các schema như star schema, tối ưu hóa cho các truy vấn phân tích và báo cáo.

Tốc độ đọc ghi: Phụ thuộc vào Cấp độ dịch vụ, kích thước cơ sở dữ liệu, lượng truy vấn thực thi song song,...

3.4.2. Mô hình triển khai



Hình 15. Hai mô hình triển khai trong Azure SQL Database. Nguồn: cloudvietnam18.com

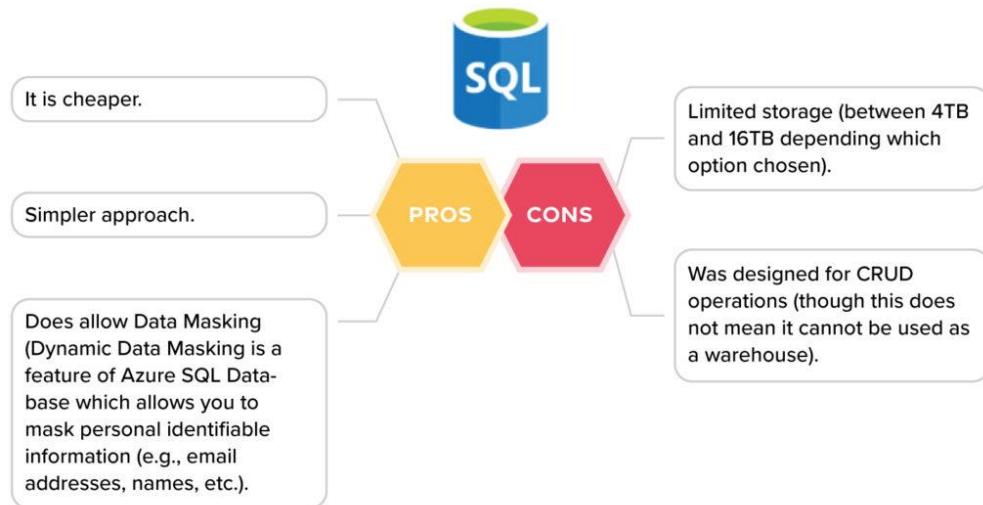
- **Single database:** Là một cơ sở dữ liệu bình thường tương tự như một cơ sở dữ liệu chứa trong công cụ cơ sở dữ liệu SQL Server. Sử dụng khi cần một người dùng duy nhất cho cloud app hoặc microservice
- **Elastic pool:** Là một tập hợp các Single database, có thể thêm 1 Single database hoặc di chuyển Single database ra khỏi Elastic pool một cách dễ dàng. Sử dụng khi cần tách biệt các database riêng, các database có thể chia sẻ với nhau về CPU hay Storage. Đồng thời có thể phân bổ tài nguyên cho mỗi database trong Elastic Pool một cách hợp lý để tiết kiệm chi phí.

3.4.3. Đặc điểm

- **Tự động điều chỉnh (tuning):** Azure SQL Database tự động điều chỉnh dựa trên các workload pattern, giúp việc duy trì hiệu suất ở quy mô lớn trở nên dễ dàng hơn.
- **Đa dạng trong Dữ liệu:** Azure SQL Database hỗ trợ và xử lý cả dữ liệu quan hệ và các cấu trúc không quan hệ, như đồ thị, JSON, không gian và XML.
- **Linh hoạt trong việc mở rộng :** Microsoft Azure SQL Database giúp điều chỉnh tài nguyên theo nhu cầu mà không cần phải cung cấp hoặc quản lý bất kỳ cơ sở hạ tầng nào.
- **Tự động backup/restore:** Azure SQL Database có sẵn chức năng back-up theo ngày hoặc theo giờ. Ta có thể back-up và restore trong nhanh chóng, không cần quan tâm tới việc lưu trữ ở đâu (có tính thêm phí lưu trữ) đảm bảo việc khắc phục sự cố hoặc thảm họa.
- **Đảm bảo uptime:** Azure SQL Database sẽ đảm bảo server luôn hoạt động (Uptime tới 99.99%), không còn lo server chết hay đĩa hỏng.
- **Monitor và Analytic:** Azure SQL Database có sẵn dashboard để monitor trạng thái của database, thời gian query, lượng đọc ghi, các query chậm.

3.4.4. Sử dụng Azure SQL Database làm Data warehouse

- Cùng với việc bao gồm các đặc điểm trên của Azure SQL Database đó là khôi lượng dữ liệu của dự án thuộc vừa và nhỏ, phù hợp để làm một data warehouse.
- Khi tận dụng Azure SQL Database làm Data warehouse sẽ giảm thiểu đáng kể chi phí dịch vụ, quản lý đơn giản hơn so với sử dụng dịch vụ thứ ba.



Hình 16. Ưu và nhược điểm việc sử dụng Azure SQL Database làm Data warehouse

3.5. Trực quan hóa dữ liệu bằng Power BI service

3.5.1. Giới thiệu

Power BI Service là một phần mềm dưới dạng dịch vụ (SaaS) thông minh dựa trên công nghệ do Microsoft, cho phép người dùng lưu trữ, chia sẻ và quản lý các báo cáo và trực quan hóa dữ liệu. Cung cấp khả năng truy cập từ bất kỳ thiết bị nào có kết nối Internet và các tính năng như lập lịch làm mới dữ liệu, chia sẻ bằng điều khiển, và tích hợp với các dịch vụ Microsoft khác như Azure và Office 365.

Tốc độ đọc ghi: Phụ thuộc vào kích thước dữ liệu, việc làm mới dữ liệu, độ phức tạp trực quan dữ liệu,.....

3.5.2. Direct Query



Hình 17. Cơ chế chế độ truy vấn - Direct Query trên PowerBI

- DirectQuery trong Power BI hoạt động bằng cách thiết lập kết nối trực tiếp với nguồn dữ liệu. Khi tạo một báo cáo trong Power BI sử dụng chế độ DirectQuery, mỗi khi báo cáo được truy cập, các truy vấn sẽ được gửi từ Power BI đến nguồn

dữ liệu để lấy dữ liệu tương ứng. Kết quả truy vấn sẽ được trả về và hiển thị trực tiếp trên báo cáo.

- Trong quá trình này, Power BI không tải toàn bộ dữ liệu vào bộ nhớ mà chỉ truy xuất dữ liệu cần thiết từ nguồn dữ liệu mỗi khi cần. Điều này cho phép Power BI làm việc với các nguồn dữ liệu có kích thước lớn, bao gồm cả hàng petabyte dữ liệu, mà không gây tốn tài nguyên bộ nhớ hệ thống.

3.5.3. Row-level security

- Row-Level Security (RLS) là chức năng quản lý và kiểm soát dữ liệu trong Power BI dựa theo mức độ cho phép tiếp cận của người sử dụng.
- RLS cho phép người quản lý của tổ chức tạo ra một chiến lược bảo vệ dữ liệu nhờ vào khả năng hạn chế truy cập của người sử dụng. Nói cách khác, RLS đảm bảo việc người dùng chỉ có thể đọc các dữ liệu mà người quản lý cho phép, điều mà không thể kiểm soát được sau khi đã công bố bộ dữ liệu lên Workspace nếu không có RLS.

3.5.4. Đặc điểm khác

- **Truy cập vào Khối lượng Dữ liệu từ Nhiều Nguồn:** Power BI có thể truy cập vào khối lượng dữ liệu không lồ từ nhiều nguồn. Nó cho phép bạn xem, phân tích và trực quan hóa khối lượng dữ liệu không lồ không thể mở trong Excel. Power BI sử dụng các thuật toán nén mạnh mẽ để nhập và lưu vào bộ nhớ cache dữ liệu trong tệp .PBIX.
- **Tương tác UI/UX:** Power BI làm cho mọi thứ trở nên trực quan hơn rất nhiều. Nó có chức năng kéo và thả dễ dàng, với các tính năng cho phép bạn sao chép tất cả định dạng trên các trực quan tương tự.
- **Chuẩn bị Dữ liệu/ Dữ liệu lớn cùng với Azure:** Sử dụng Power BI với Azure cho phép bạn phân tích và chia sẻ khối lượng dữ liệu không lồ. Ví dụ Azure Data Lake giúp giảm thời gian trao đổi những hiểu biết và tăng cường hợp tác giữa các người dùng.

- **Phân tích luồng thời gian thực:** Power BI giúp bạn lấy dữ liệu từ nhiều cảm biến và nguồn truyền thông xã hội để có quyền truy cập vào phân tích thời gian thực, giúp bạn luôn sẵn sàng đưa ra các quyết định.

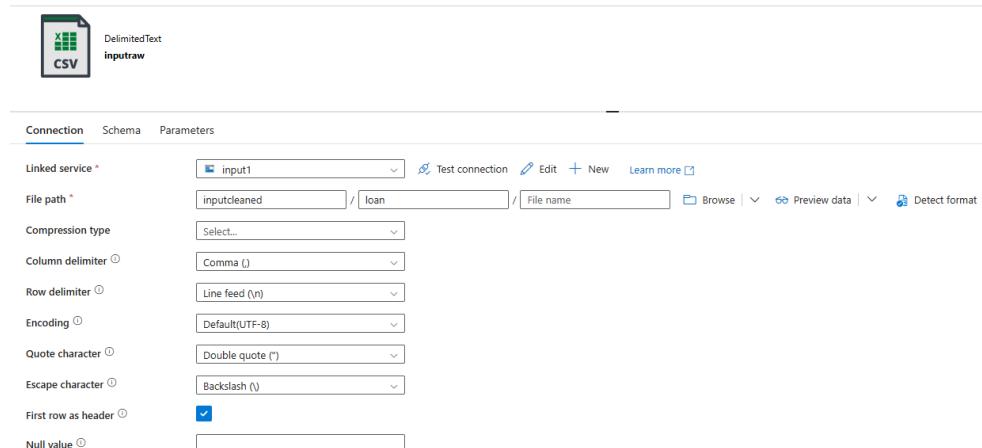
III. MÔ HÌNH DỮ LIỆU - TRIỂN KHAI

1. Kết quả triển khai mô hình

1.1. Luồng xử lý dữ liệu tự động ETL

1.1.1. Data pipeline

- **Extract:** Dữ liệu được trích xuất từ Azure Blob Storage, thông qua linked service cung cấp từ Azure Data Factory.



Hình 18. Cấu hình linked service - dataset với nguồn dữ liệu Azure Blob Storage

- **Transform - Load:** Dữ liệu được biến đổi theo các kỹ thuật được đề cập trước đó và Load vào datawarehouse house theo mô hình Star schema.



Hình 19. Tổng quan quy trình Transform trong Azure Data Factory

```

1
import numpy as np
from pyspark.sql import SparkSession
from pyspark.ml.feature import Imputer
from pyspark.sql.types import FloatType, ArrayType
from pyspark.sql.functions import col, to_date, year, month, regexp_like,
    regexp_extract, isnan, when, count, lit, udf, \
    pow as psf_pow, sum as psf_sum
import findspark
findspark.init()

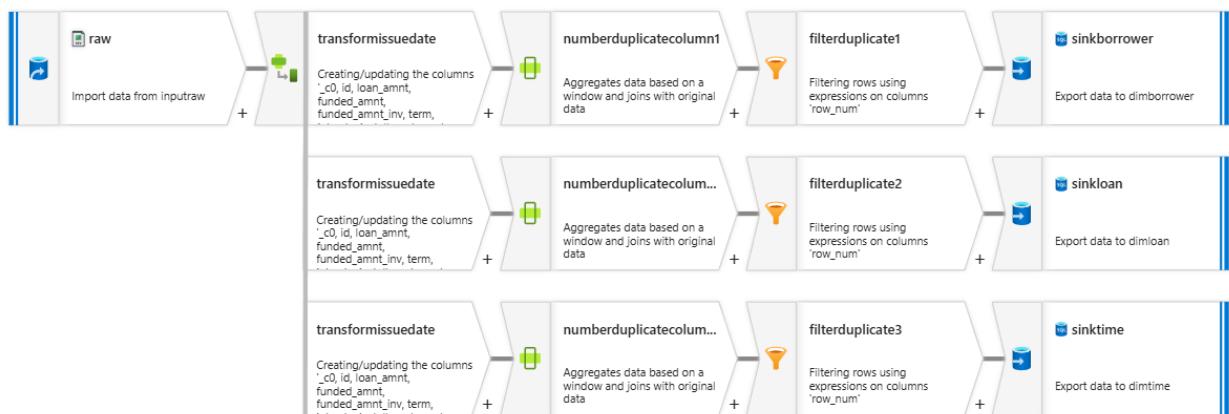
2
import warnings
warnings.filterwarnings('ignore')

3
storage_account_name = 'storageaccfinalproject'
storage_account_access_key = '0UIIugAE05nU1sqavDh4fHMy/CehNs6GVCI#x8dnbujqRCYGSzRovY2nUZ2g6v18016Nen211IyD+ASt/gUMmQ=='
spark.conf.set('fs.azure.account.key.' + storage_account_name +
    '.blob.core.windows.net', storage_account_access_key)

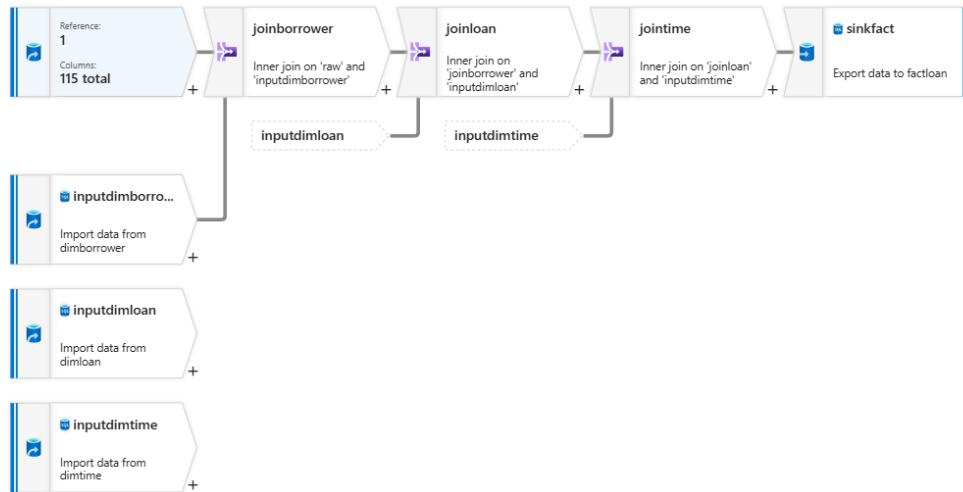
4
blob_container = 'inputraw'
filePath = "wasbs://" + blob_container + "g" + storage_account_name + \
    ".blob.core.windows.net/loan_status_2007-2009.csv"
df = spark.read.format("csv").load(filePath, inferSchema=True, header=True)

```

Hình 20. Chi tiết code xem qua file notebook *data_pre_processing.ipynb*



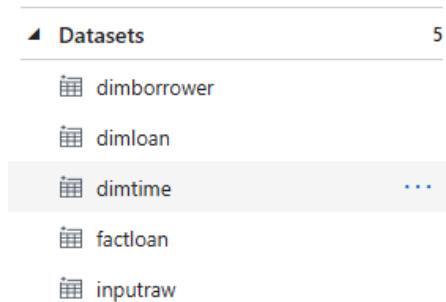
Hình 21. Tóm tắt cách xử lý dữ liệu thành các bảng Dimensions (Mapping Data flow)



Hình 22. Tạo bảng Fact từ các bảng Dim và dữ liệu ban đầu (Mapping Data flow)

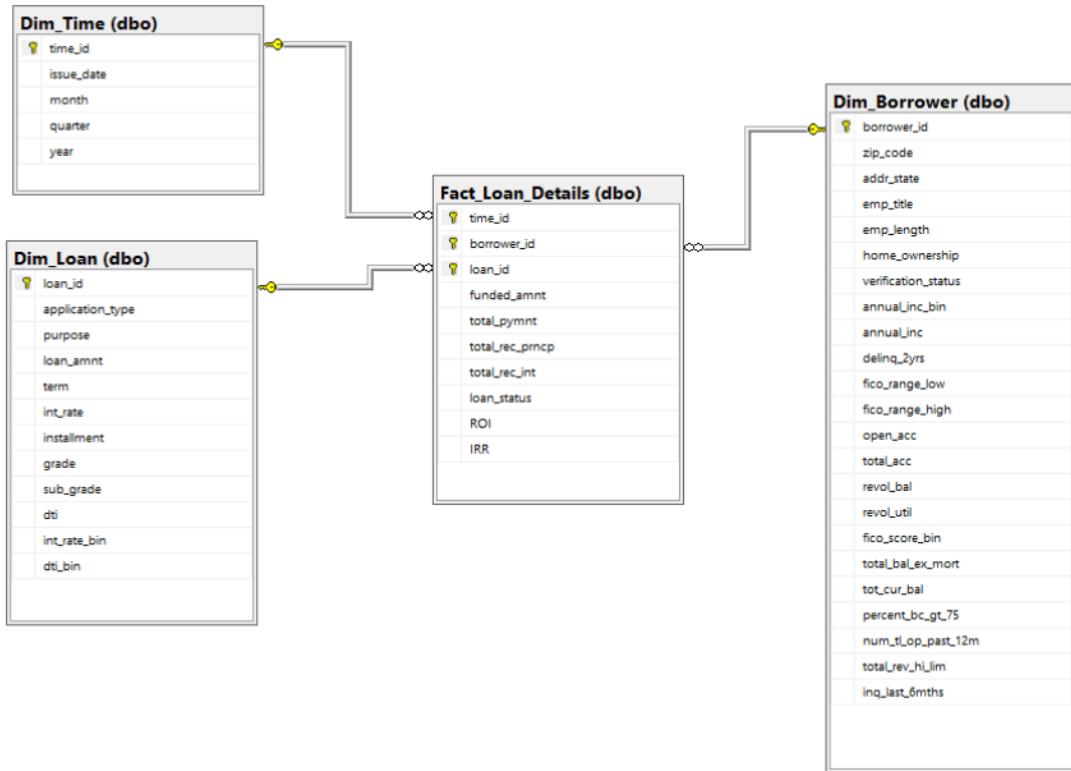
Linked Service	Connection	Schema	Parameters
dimborrower	Linked service * inputsqli	dbo.Dim_Borrower	<input type="checkbox"/> Enter manually
dimloan	Linked service * inputsqli	dbo.Dim_Loan	<input type="checkbox"/> Enter manually
dimborrower	Linked service * inputsqli	dbo.Dim_Borrower	<input type="checkbox"/> Enter manually
dimloan	Linked service * inputsqli	dbo.Dim_Loan	<input type="checkbox"/> Enter manually

Hình 23. Cấu hình Linked service đến các bảng trong Azure SQL Database



Hình 24. Các datasets kết nối với các Linked services

1.1.2. Định dạng lưu trữ



Hình 25. Mô hình Star schema trong Data warehouse sau quá trình ETL

- Thông tin các bảng dimension (chiều) và fact (sự kiện) trong data warehouse
 - + **Dim_Time**: 93 dòng x 5 cột
 - + **Dim_Loan**: 1,739,918 dòng x 12 cột
 - + **Dim_Borrower**: 1,767,177 dòng x 23 cột
 - + **Fact_Loan_Details**: 1,767,250 dòng x 10 cột

1.1.3. Độ tốc độ thực thi

Activity name ↑↓	Activity status ↑↓	Activity type ↑↓	Run start ↑↓	Duration ↑↓	Integration runtime
import_fact_table	✔ Succeeded	Data flow	5/13/2024, 8:57:04 PM	2m 46s	AutoResolveIntegration
import_dim_table	✔ Succeeded	Data flow	5/13/2024, 8:48:49 PM	8m 14s	AutoResolveIntegration
Re-create table SQL	✔ Succeeded	Script	5/13/2024, 8:48:35 PM	13s	AutoResolveIntegration
Tranform data	✔ Succeeded	Notebook	5/13/2024, 8:40:25 PM	8m 7s	AutoResolveIntegration

Hình 26. Tốc độ thực thi của các thành phần trong Data pipeline

- Thành phần Azure Databricks mất trung bình ~8 phút cho tác vụ xử lý dữ liệu

- Thành phần Mapping Dataflow mất trung bình ~10 phút cho tác vụ tổ chức dữ liệu

1.1.4. Đo độ trễ khi thiết lập server ở 2 vùng khác nhau

The image shows two separate Power BI dashboards. The top dashboard is titled "Southeast Asia" and the bottom one is titled "East US". Both dashboards display a table of activity runtimes. The columns include Activity name, Activity status, Activity type, Run start, Duration, and Integration runtime.

Activity name	Activity status	Activity type	Run start	Duration	Integration runtime
import_fact_table	Succeeded	Data flow	5/13/2024, 8:57:04 PM	2m 46s	AutoResolveIntegration
import_dim_table	Succeeded	Data flow	5/13/2024, 8:48:49 PM	8m 14s	AutoResolveIntegration
Re-create table SQL	Succeeded	Script	5/13/2024, 8:48:35 PM	13s	AutoResolveIntegration
Transform data	Succeeded	Notebook	5/13/2024, 8:40:25 PM	8m 7s	AutoResolveIntegration

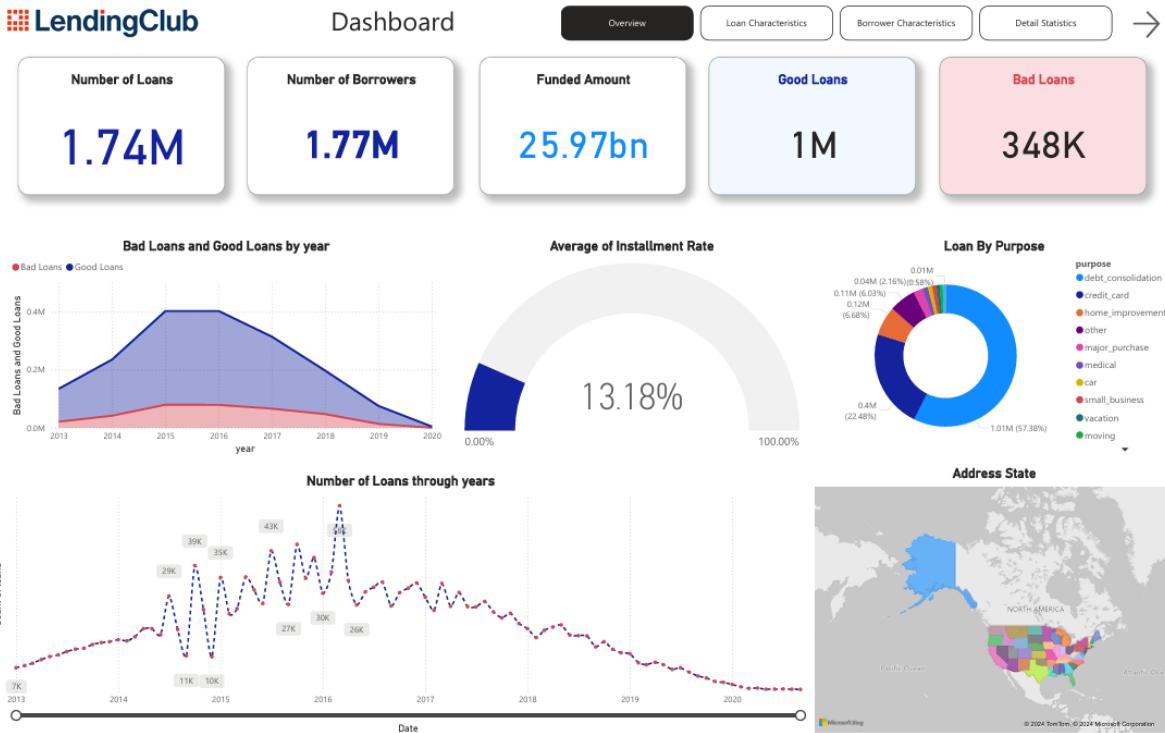
Activity name	Activity status	Activity type	Run start	Duration	Integration runtime
import_fact_table	Succeeded	Data flow	5/13/2024, 9:43:07 PM	1m 58s	AutoResolveIntegration
import_dim_table	Succeeded	Data flow	5/13/2024, 9:35:21 PM	7m 45s	AutoResolveIntegration
Re-create table SQL	Succeeded	Script	5/13/2024, 9:34:26 PM	54s	AutoResolveIntegration
Transform data	Succeeded	Notebook	5/13/2024, 9:27:19 PM	7m 7s	AutoResolveIntegration

Hình 27. Thực hiện đo độ trễ giữa 2 server ở 2 region khác nhau

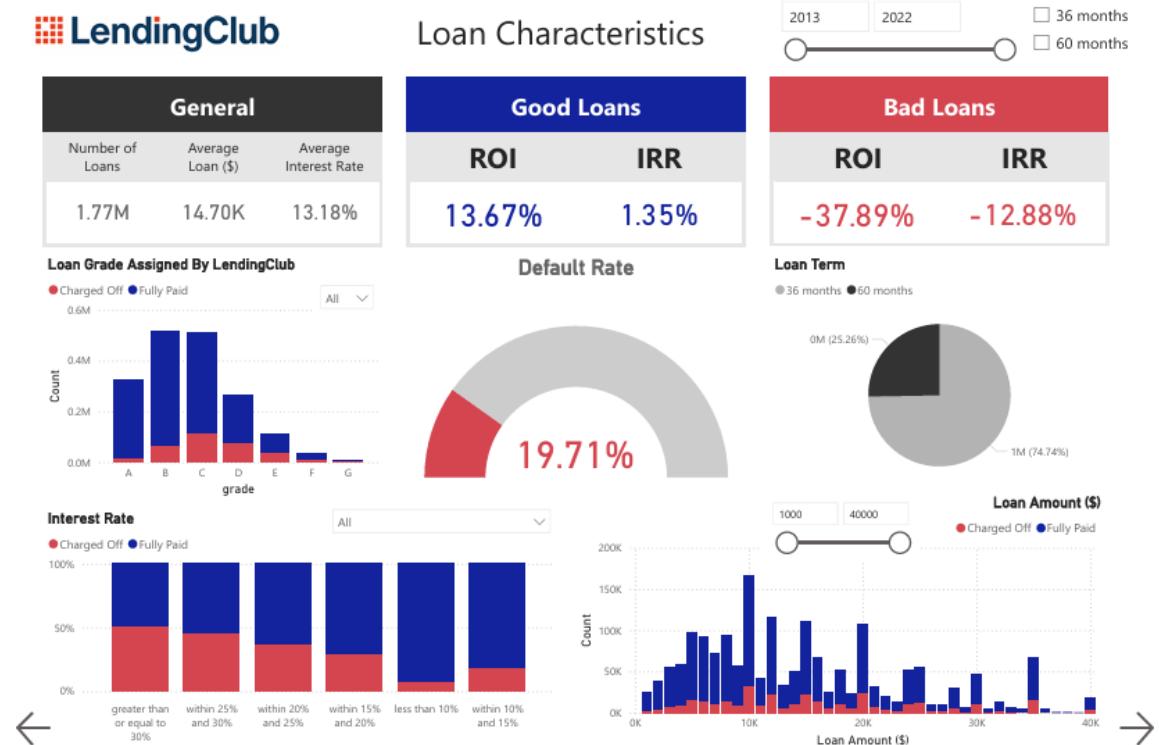
- Thiết lập 2 server cùng thao tác các tác vụ của luồng xử lý dữ liệu tự động ETL, bao gồm Southeast Asia (SEA) và East US
- Nhận xét:
 - + Thời gian của server ở khu vực East US có một chút nhanh hơn đối với SEA với đa số tác vụ
 - + Tuy nhiên có sự khác biệt rõ ràng ở tác vụ chạy Script tạo table đối với EastUS và SEA, EastUS có sự chậm hơn đáng kể đối với SEA.

1.2. Hiện thực trực quan hóa dữ liệu

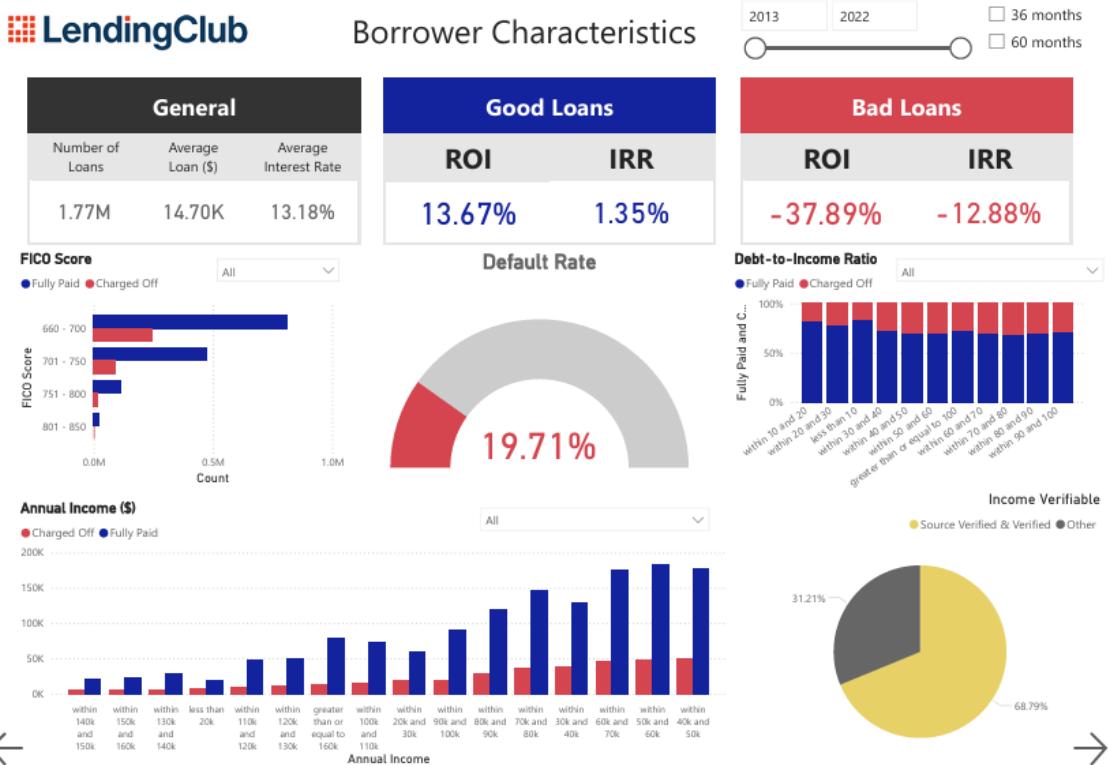
1.2.1. Tổng quan dashboard



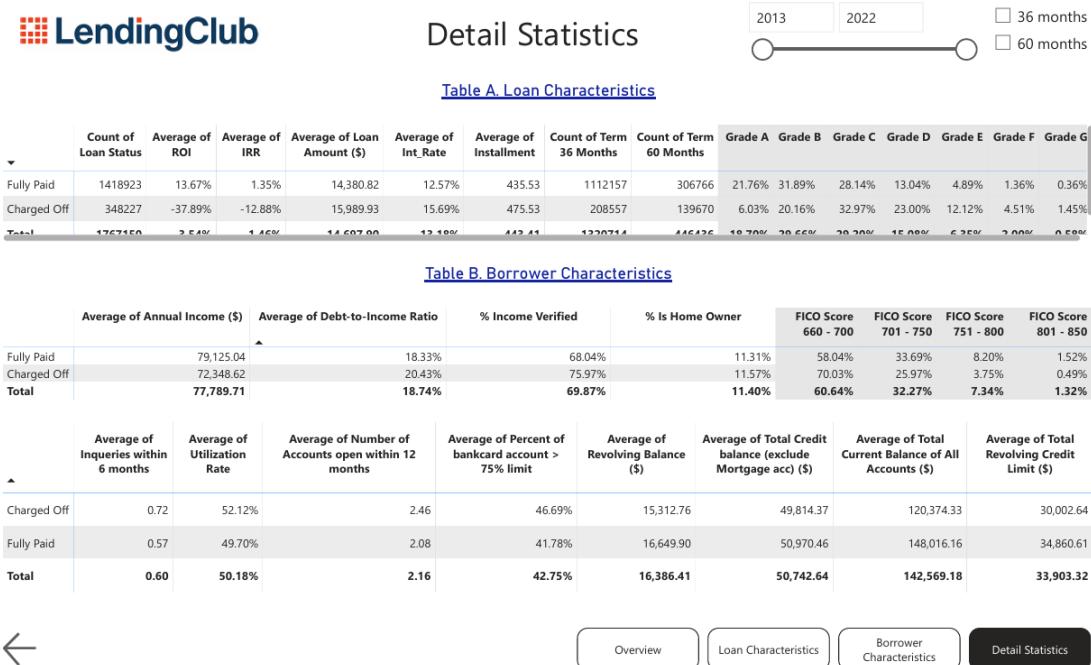
Hình 28. Trang tổng quan.



Hình 29. Trang filter 1: ĐẶC TRƯNG VỀ CÁC KHOẢN VAY



Hình 30. Trang filter 2: Đặc trưng về người vay



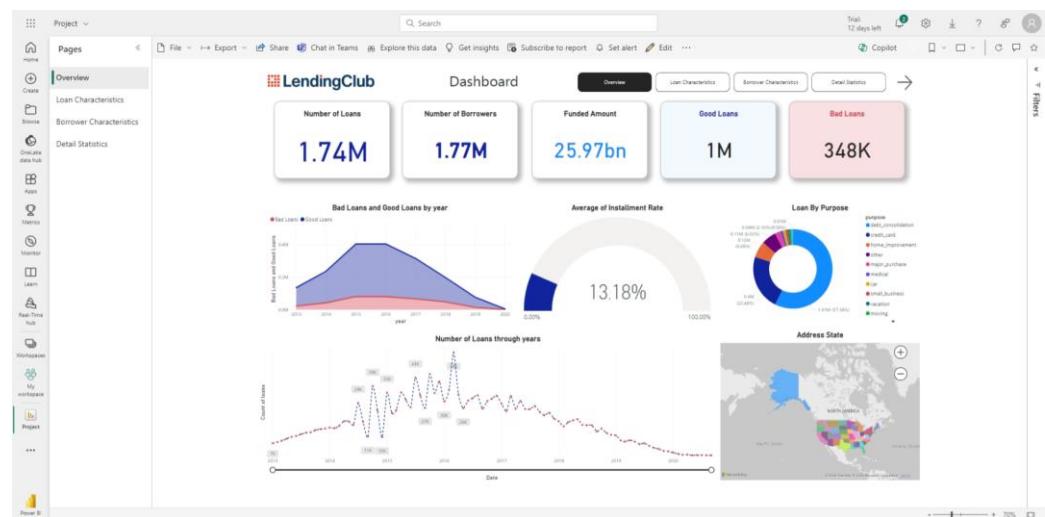
Hình 31. Trang chi tiết về các thông số của công ty

1.2.2. Các chức năng chi tiết đã hiện thực trên dashboard

- **Deploy dashboard lên PowerBI service:** Cấu hình publish dashboard từ PowerBI Desktop lên PowerBI service.



Hình 32. Xuất bản lên PowerBI Service



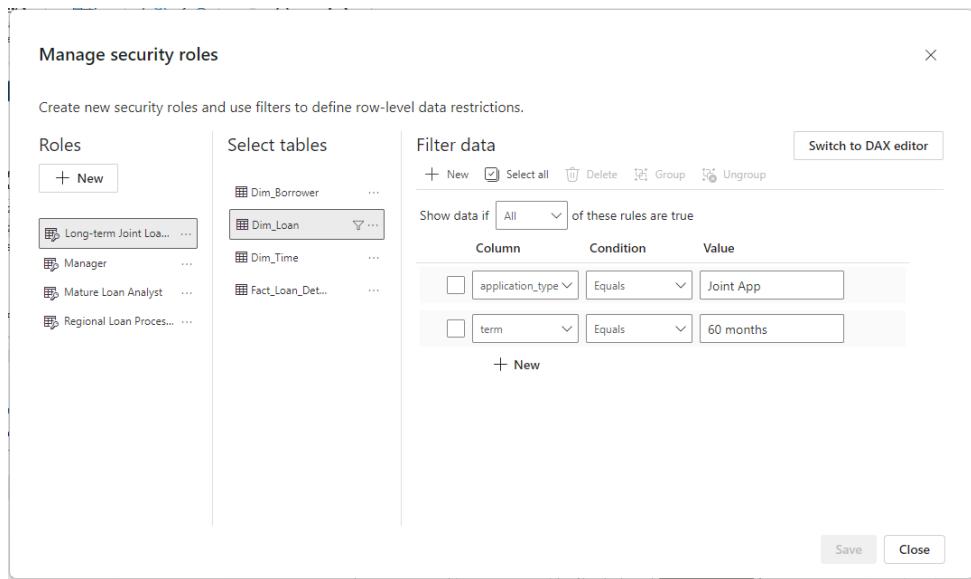
Hình 33. Giao diện trên PowerBI service

- **Triển khai cập nhật dữ liệu tức thời từ mô hình dữ liệu lên dashboard:** Sử dụng chế độ kết nối Direct Query để tạo live-connection với Azure SQL Database



Hình 34. Kết nối với mô hình dữ liệu sử dụng Direct Query

- **Bảo mật chia sẻ và ràng buộc điều kiện dữ liệu trên dashboard đến người dùng:** Sử dụng cơ chế Row-level Security để phân chia các vai trò của người dùng truy cập vào dashboard
 - + Các vai trò và điều kiện như sau:
 - Quản lý (Manager): truy cập toàn bộ không giới hạn
 - Chuyên viên phân tích các khoản vay doanh nghiệp – dài hạn (Long-term Joint Loan Analyst):
 - (Dim_Loan) [application_type] == "Joint App" && [term] == "60 months"
 - (Dim_Loan) [grade] IN {"A", "B", "C"}
 - (Dim_Time) [year] > 2015
 - Người xử lý các người vay theo vùng miền (Regional Loan Processor):
 - (Dim_Borrower)
 - [addr_state] IN {"PA", "KY", "IN", "HI", "VA", "KS", "DE", "SC", "DC", "NH"}



Hình 35. Cấu hình điều kiện cho các vai trò người dùng trong giao diện desktop

Row-Level Security

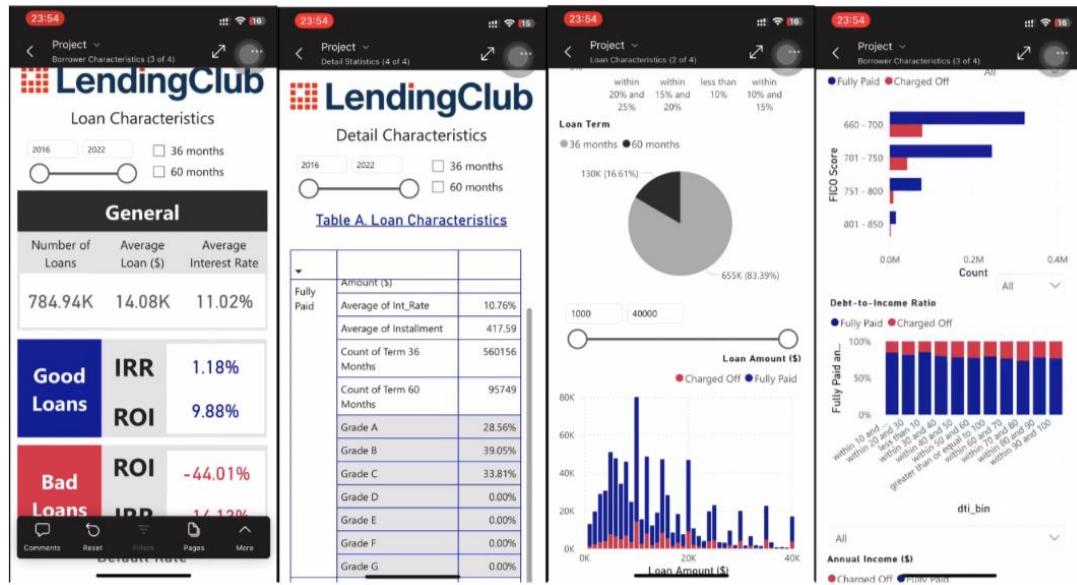
Long-term Joint Loan Analyst (0) Manager (1) Mature Loan Analyst (1) Regional Loan Processor (0)	Members (1) People or groups who belong to this role <input type="text" value="Enter email addresses"/> <input type="button" value="Add"/> Lê Quốc Khanh ×
---	--

Row-Level Security

Long-term Joint Loan Analyst (0) Manager (1) Mature Loan Analyst (1) Regional Loan Processor (0)	Members (1) People or groups who belong to this role <input type="text" value="Enter email addresses"/> <input type="button" value="Add"/> Lê Gia Kiệt ×
--	--

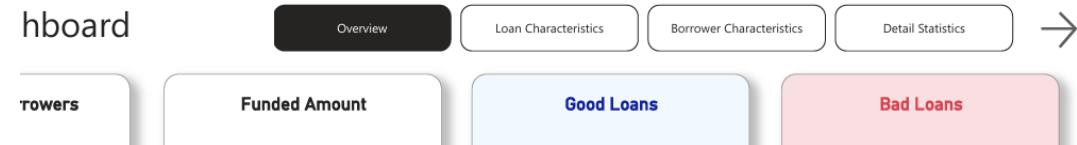
Hình 36. Chỉ định vai trò cho các tài khoản người dùng truy cập

- **Thực hiện Responsive trên thiết bị di động:** Tạo các giao diện responsive cho các người dùng truy cập bằng điện thoại, tablet,..

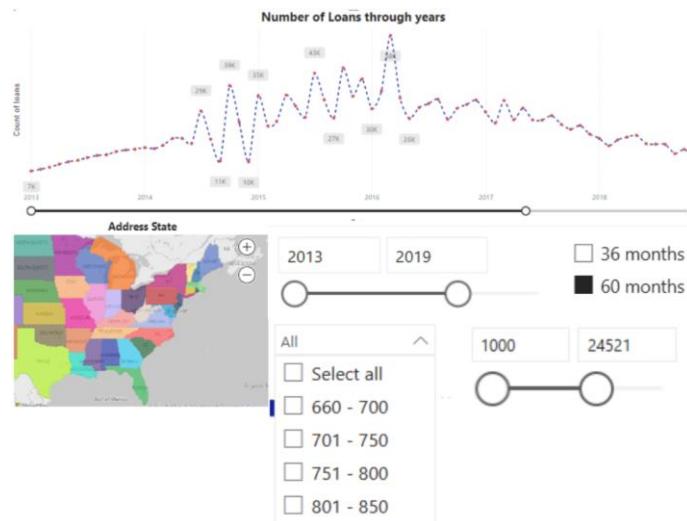


Hình 37. Một số giao diện dashboard trên mobile

- **Một số thành phần giúp tăng tương tác với dashboard:** Sử dụng các button, navigation để điều hướng giữa các report và các thành phần tương tác (slicer) khác như slider, combo-box,..



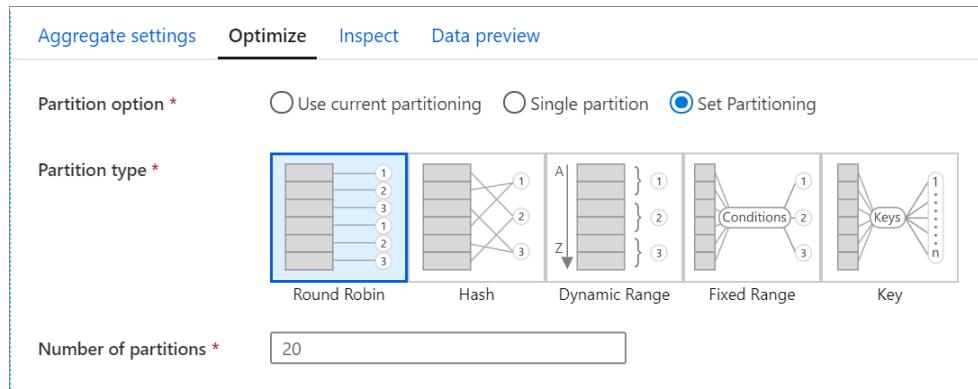
Hình 38. Thành phần điều hướng giữa các trang report



Hình 39. Slicer cho report với các thành phần như slider, combo-box, map

1.3. Giải pháp tối ưu luồng xử lý dữ liệu

- Tab Optimize trong Data Flow của Azure Data Factory chứa các thiết lập để cấu hình sơ đồ phân vùng của cluster. Việc điều chỉnh phân vùng cung cấp quyền kiểm soát phân phối dữ liệu trên các node và tối ưu hóa vị trí dữ liệu, hiệu suất tổng thể của luồng dữ liệu.
- Kiểu phân vùng (Partition Type):



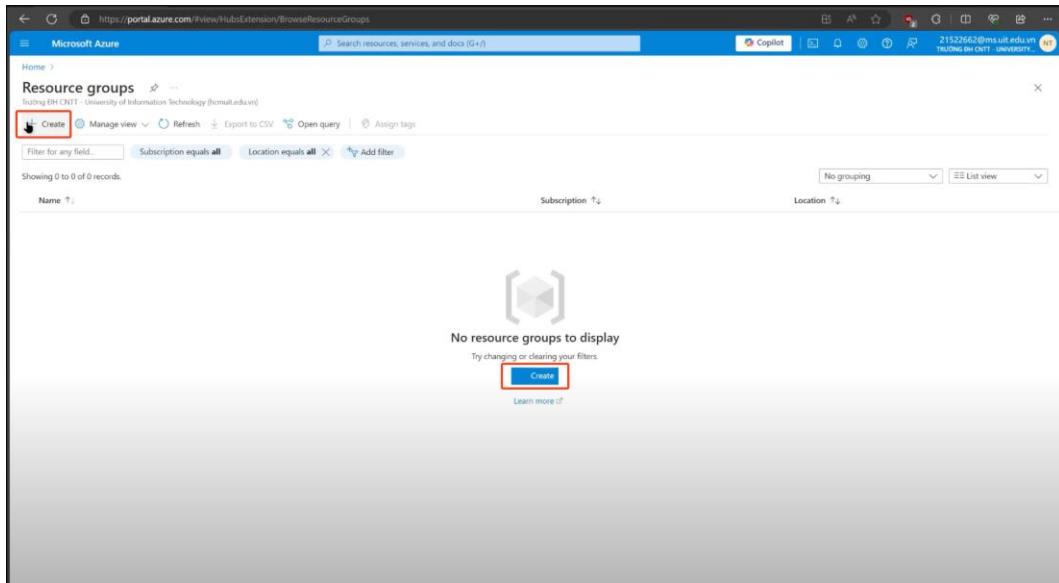
Hình 40. Các loại Phân vùng cho nguồn dữ liệu. Nguồn: learn.microsoft.com

- + Round Robin: Phân phối dữ liệu đồng đều giữa các worker node.
- + Hash: Phân vùng dữ liệu dựa trên giá trị băm của một hoặc nhiều cột. Sử dụng khi cần phân phối dữ liệu đồng đều theo giá trị.
- + Dynamic Range: Phân vùng dữ liệu dựa trên phạm vi giá trị của một cột. Có thể gây lệch phân vùng
- + Fixed Range: Phân vùng dữ liệu dựa trên phạm vi giá trị cố định như phân vùng theo ngày tháng
- + Key: Phân vùng dữ liệu dựa trên giá trị của một cột khóa. Nhưng khó kiểm soát số lượng phân vùng.

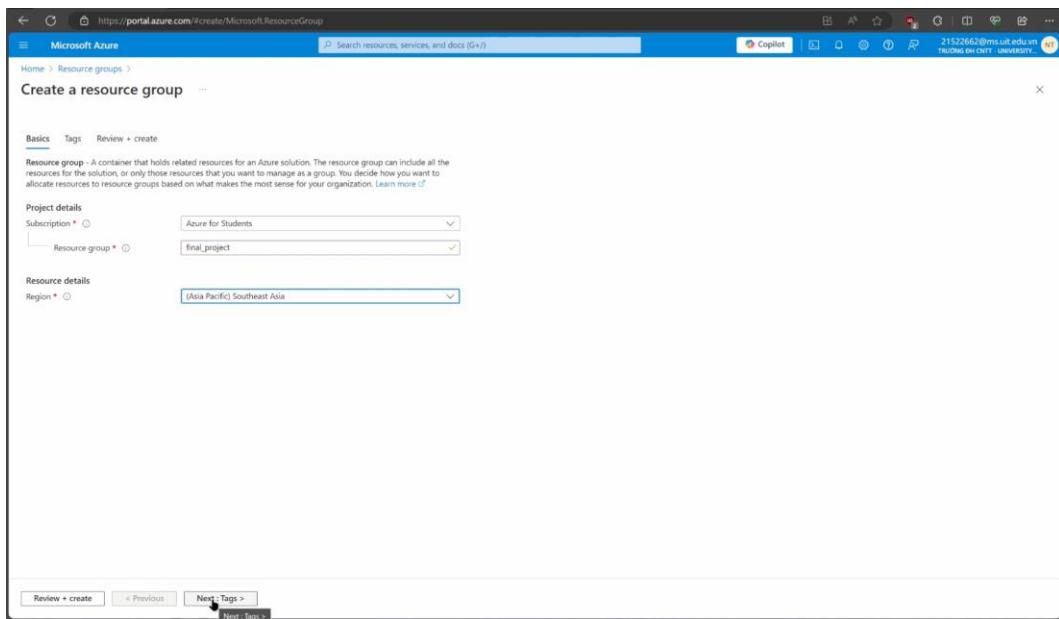
2. Chi tiết hiện thực triển khai

2.1. Luồng xử lý dữ liệu tự động ETL

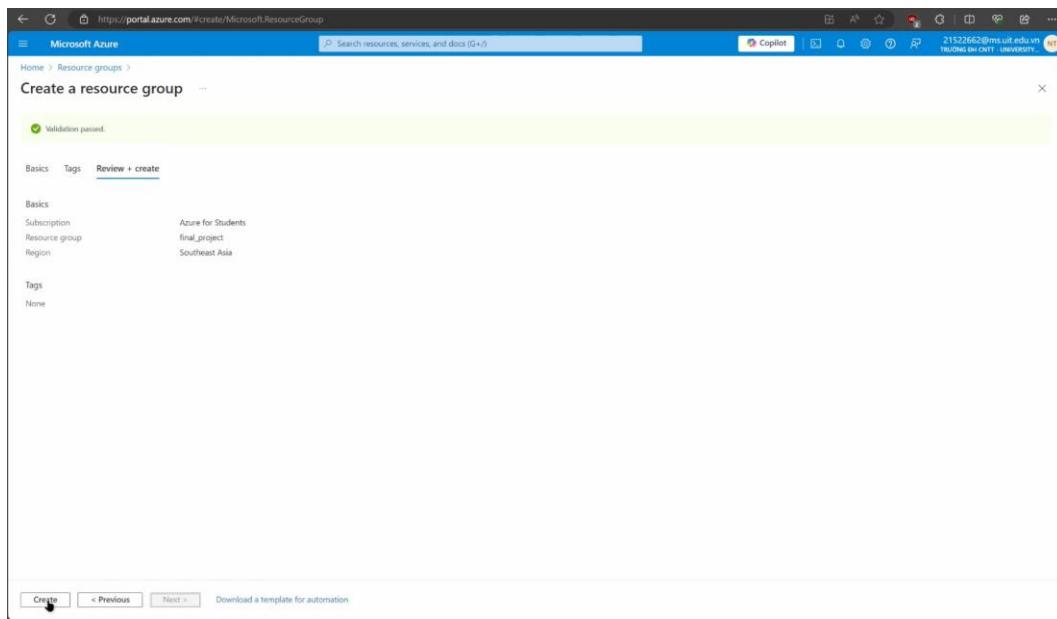
2.1.1. Cài đặt Resource Group



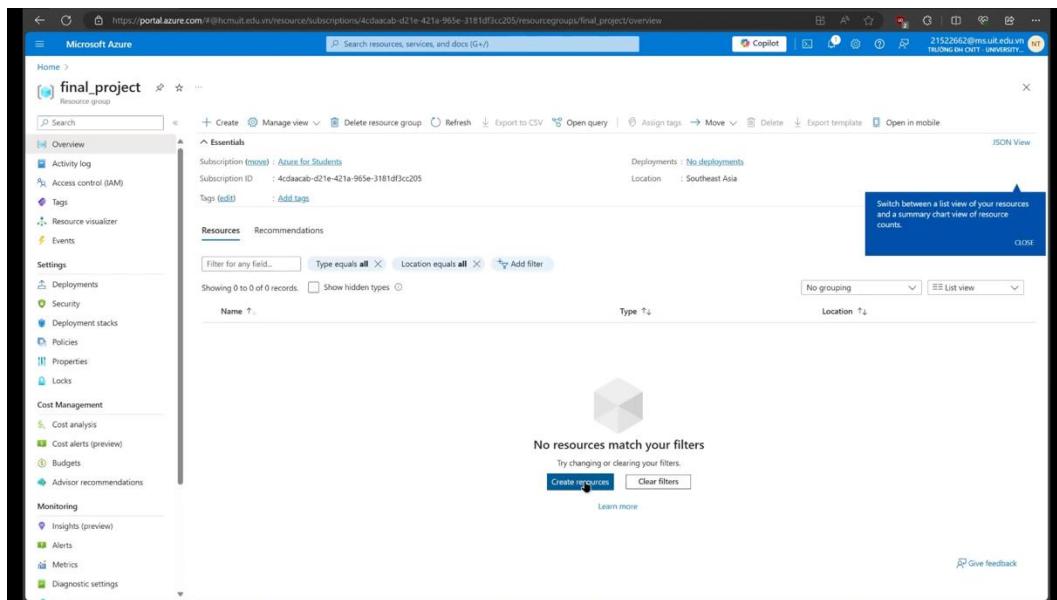
Hình 41. Trong tab Resource groups, chọn Create để tạo mới một Resource group



Hình 42. Nhập các thông tin về tên, khu vực và tùy chọn gói tại tab Basics

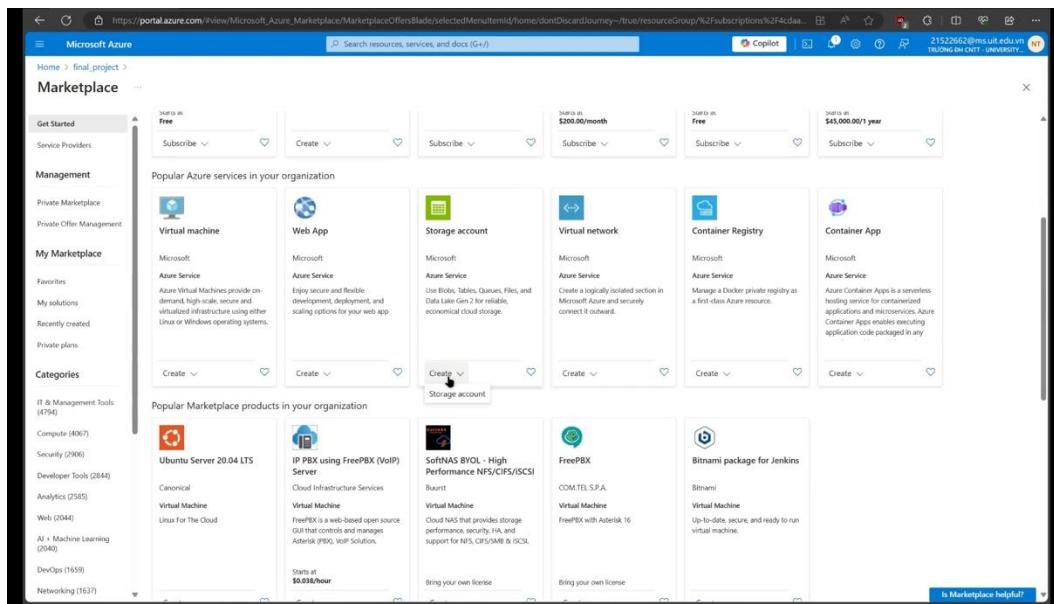


Hình 43. Tại tab Review + create, sau khi Azure review các thông tin vừa nhập thì chọn Create để hoàn tất quá trình tạo Resource group

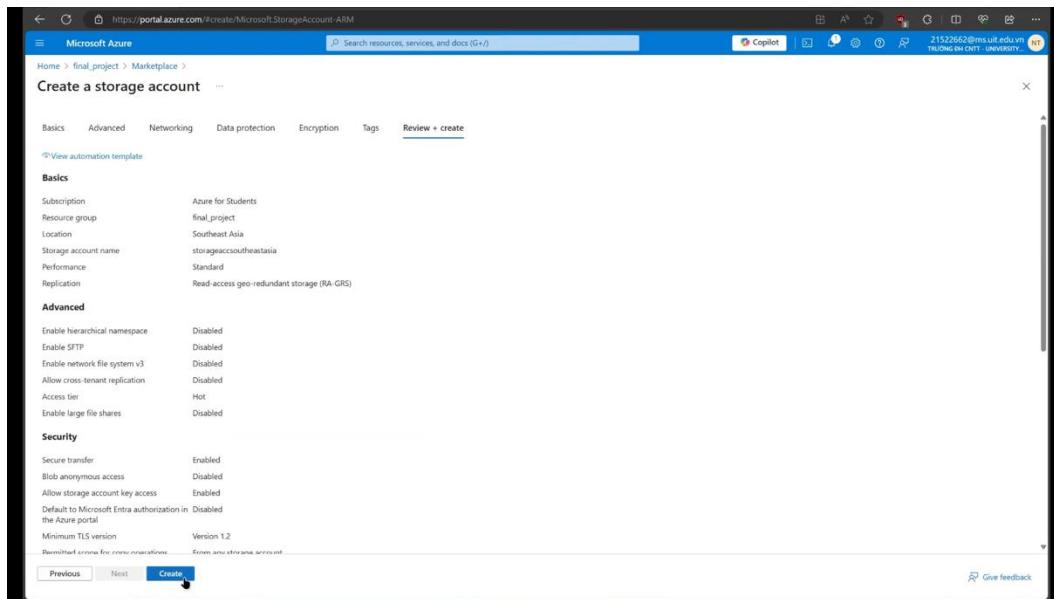


Hình 44. Chọn Create Resources để tạo những Resource cần trong Resource group

2.1.2. Cài đặt Storage Account



Hình 45. Tìm kiếm và chọn Storage account trong Marketplace

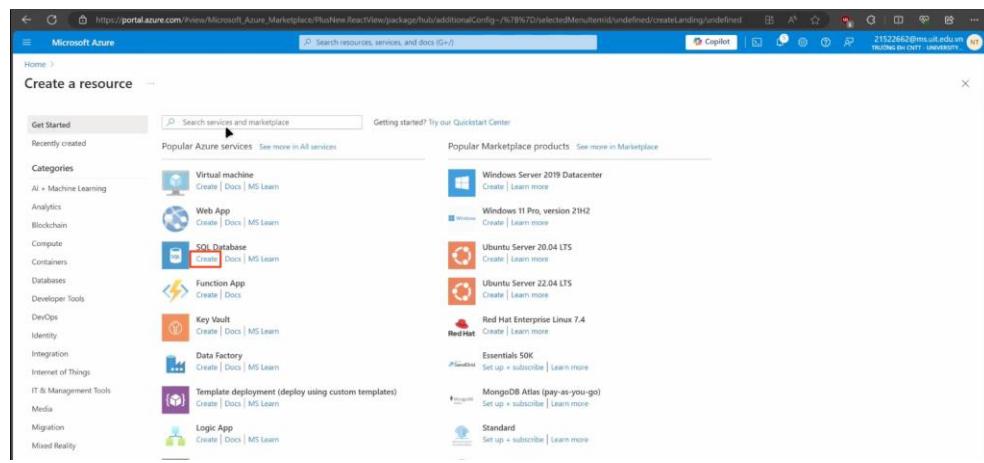


Hình 46. Sau khi Azure review lại thông tin vừa nhập ở các tab thì chọn Create để hoàn tất tạo Storage Account

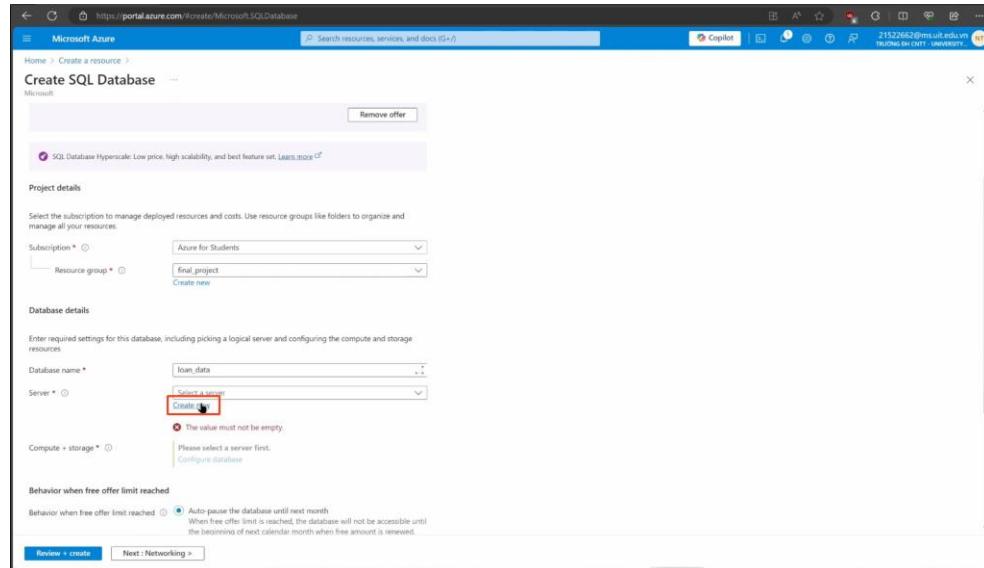
Hình 47. Trong phần Data storage, chọn Containers và nhấn + Container để tạo mới

Hình 48. Tạo mới hai Container là "inputraw" và "inputclean" và upload các file csv vào

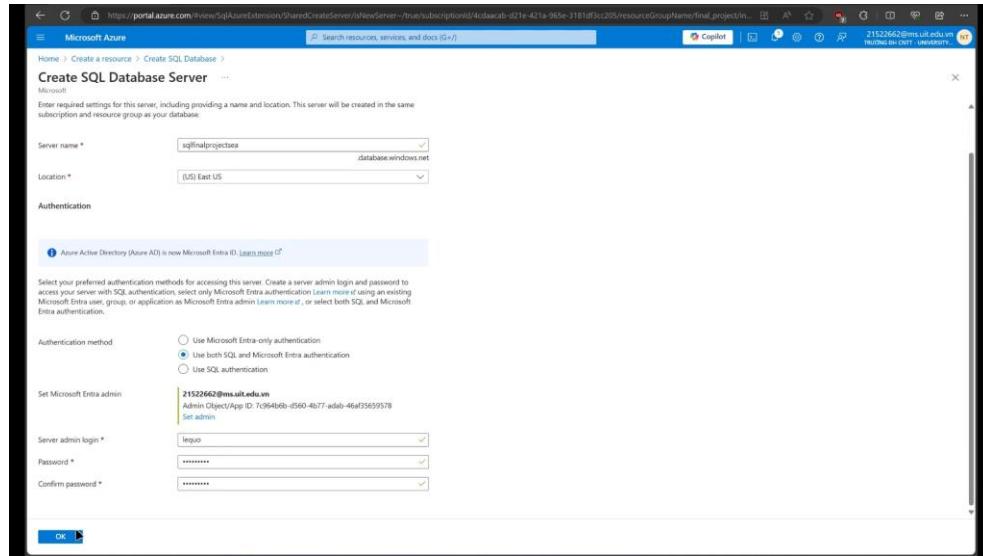
2.1.3. Cài đặt SQL Database and SQL server



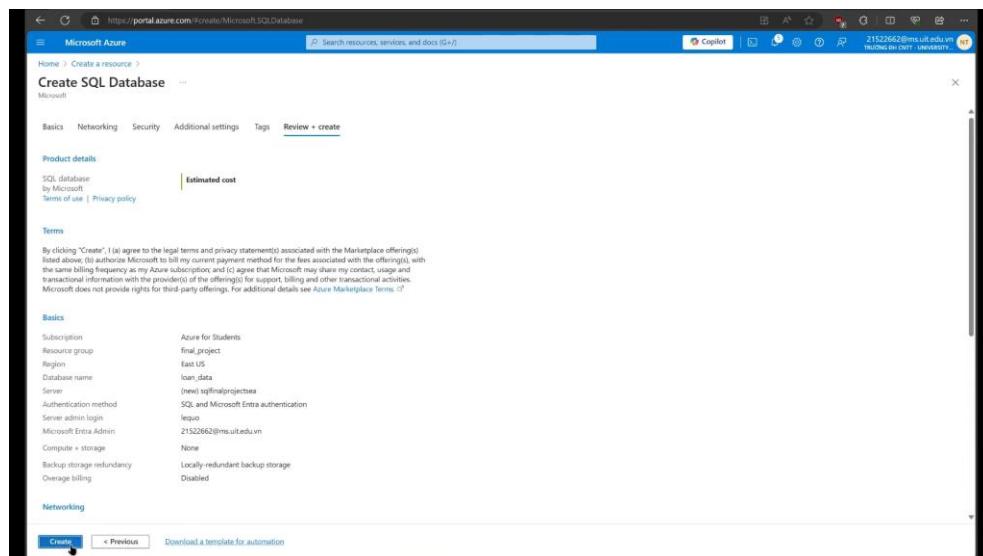
Hình 49. Vào Marketplace chọn Create SQL Database



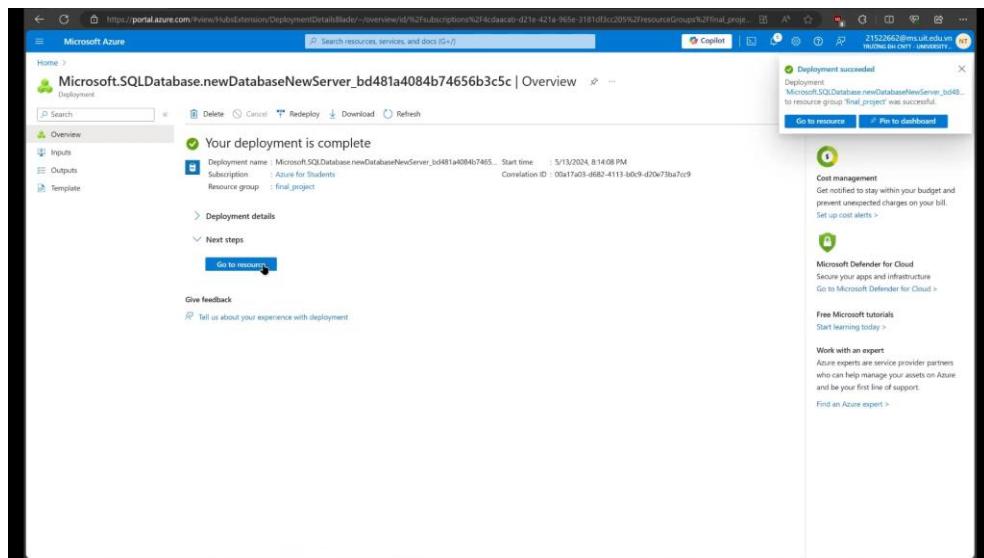
Hình 50. Nhập các thông tin và chọn Create new ở mục Server



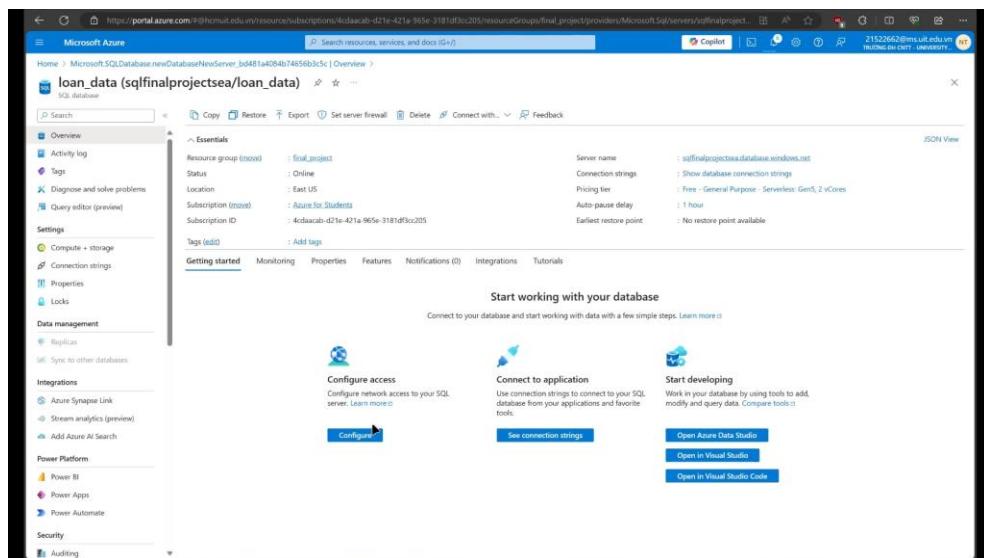
Hình 51. Nhập các thông tin và nhấn *Create* để tạo mới Server



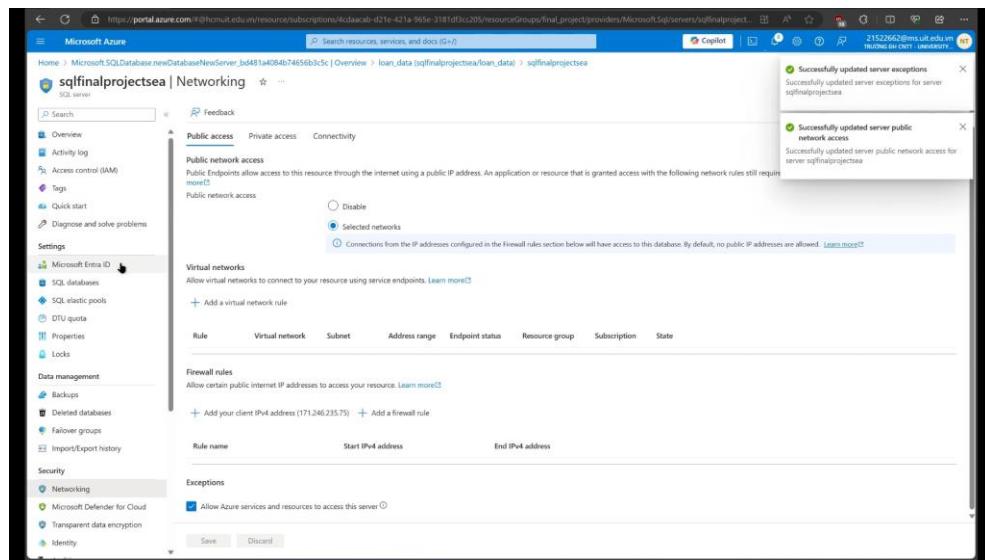
Hình 52. Sau đó, khi hoàn thành các tab thì chọn *Create* để tạo mới SQL Database



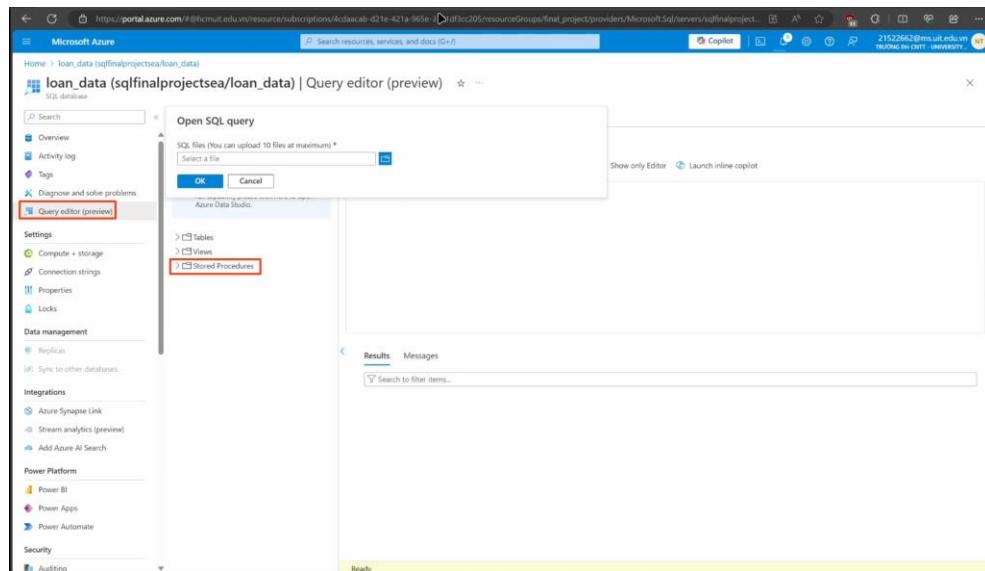
Hình 53. Tiếp đó đợi Azure deploy thành công



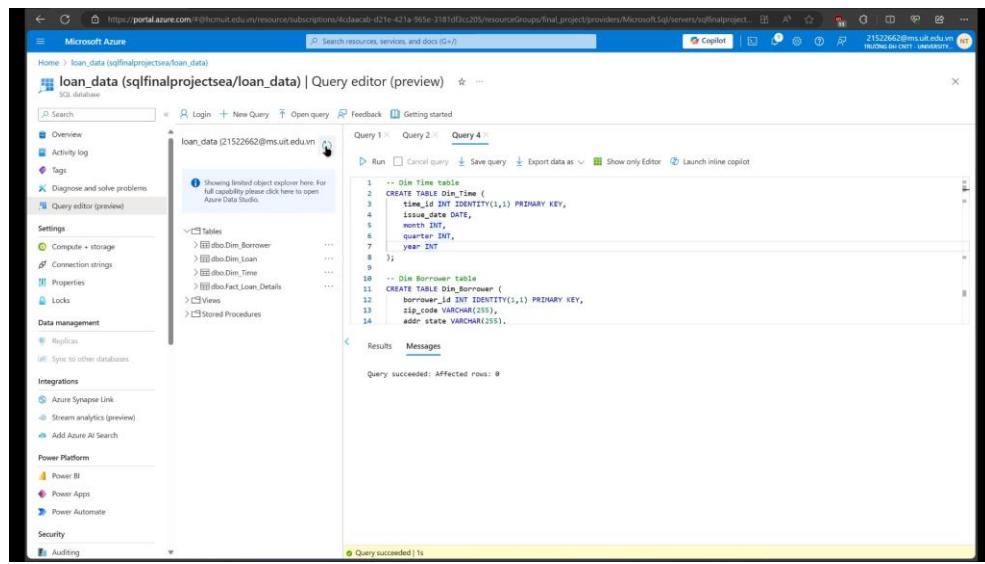
Hình 54. Tiếp tục chọn *Configure* ở mục *Configure access* để cấu hình truy cập cho SQL



Hình 55. Cập nhật lại cấu hình như trên

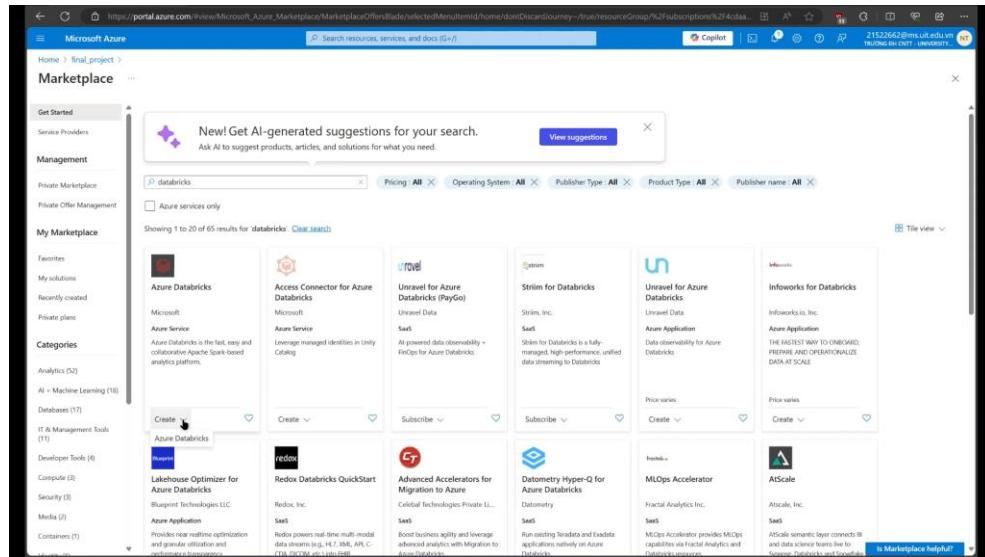


Hình 56. Vào lại file SQL Database loan_data, chọn Query editor và Stored Procedures để upload truy vấn

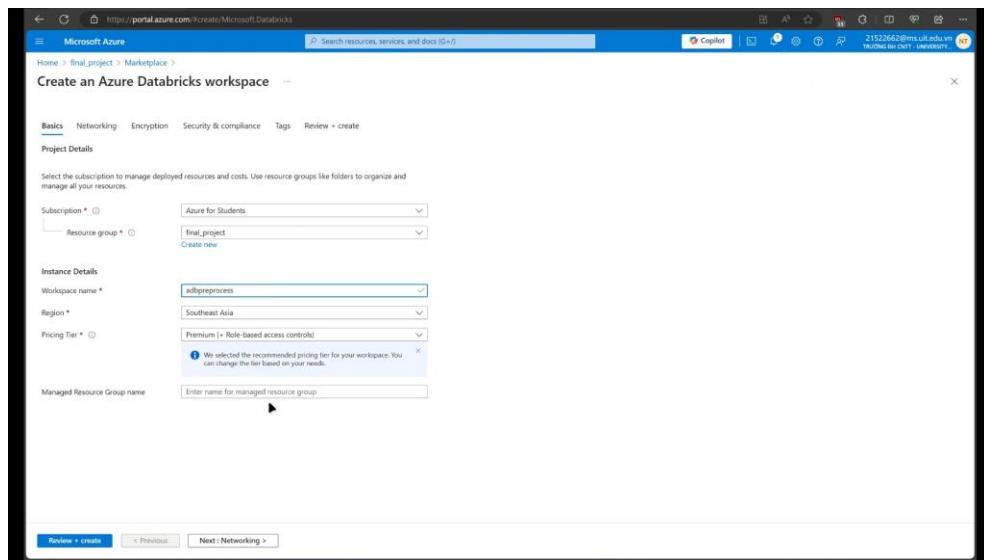


Hình 57. Sau đó chọn Run để chạy các câu truy vấn

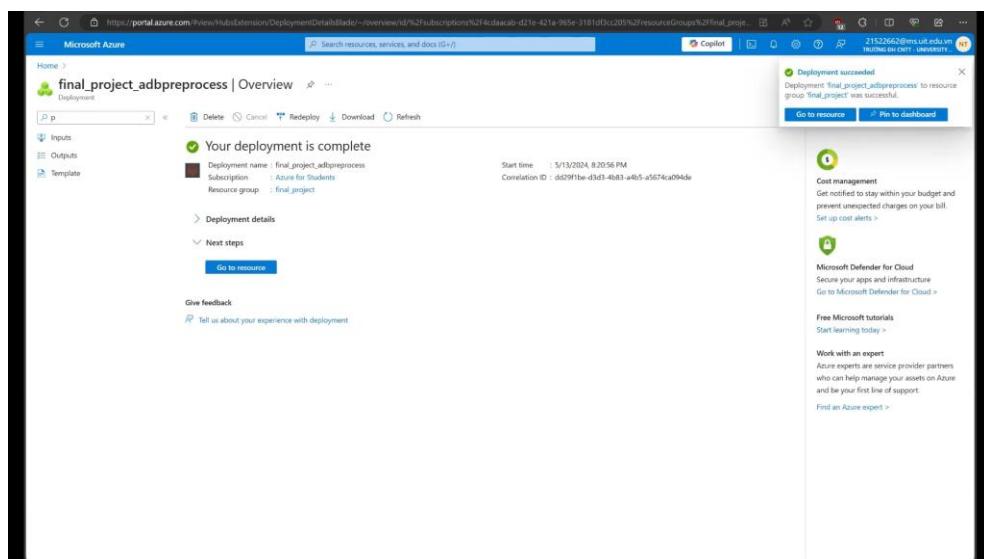
2.1.4. Cài đặt Azure Databricks



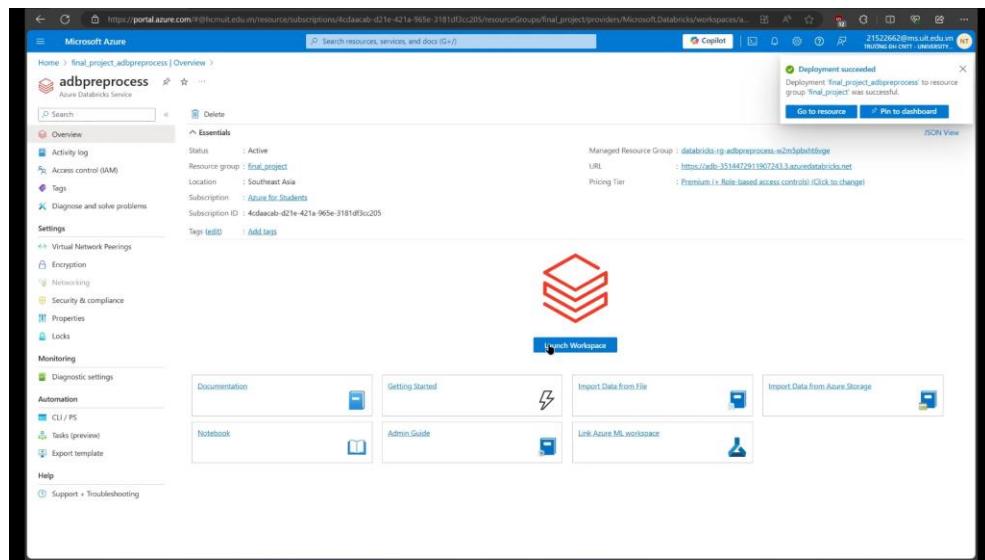
Hình 58. Quay lại Marketplace tìm Databricks chọn Create để tạo mới



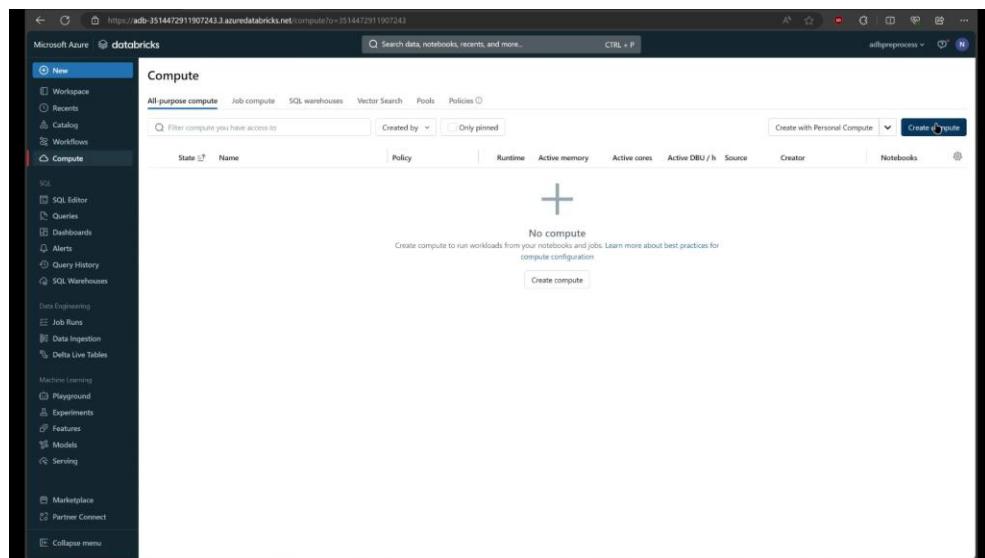
Hình 59. Nhập thông tin các tab, chọn Review + create và đợi Azure review xong thì chọn Create để tạo mới



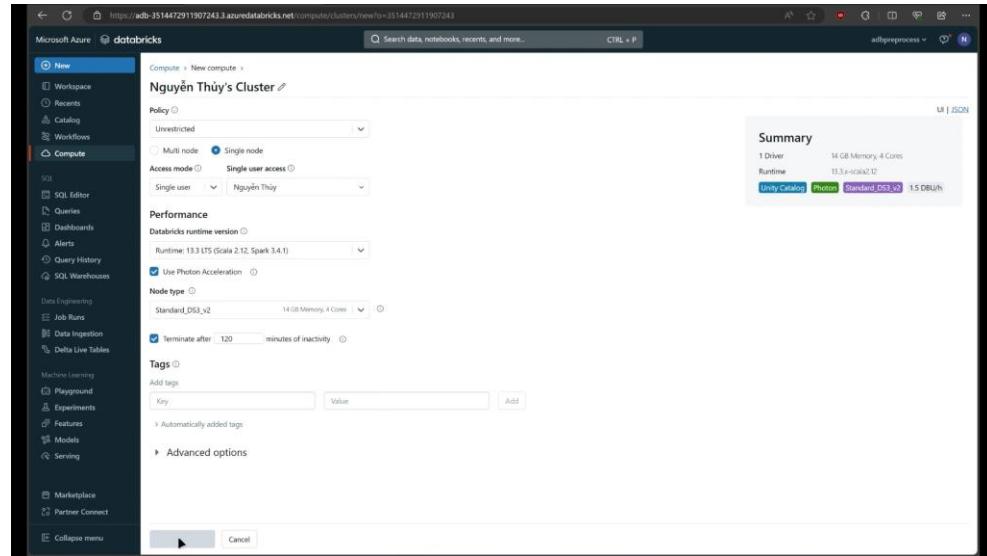
Hình 60. Deploy thành công



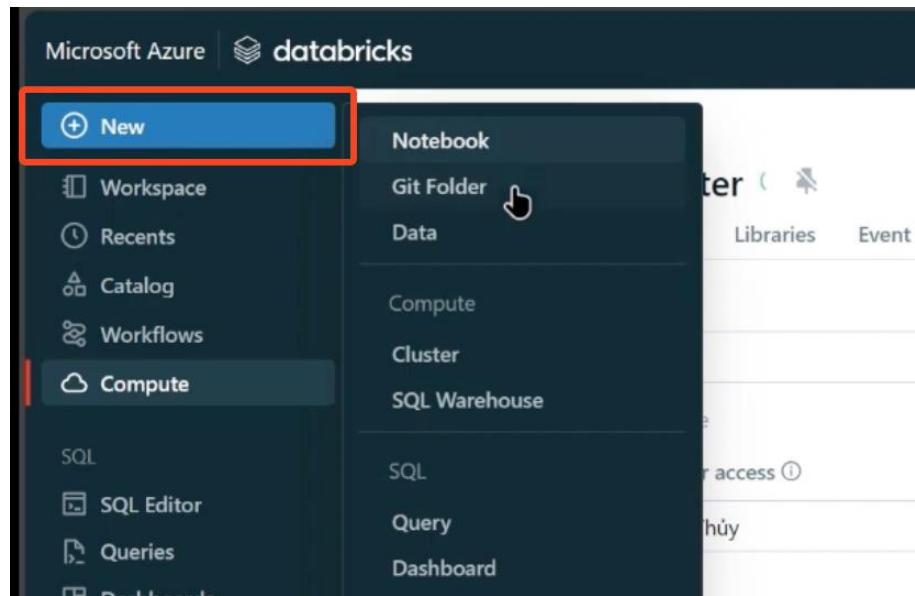
Hình 61. Tại đây chọn Launch Workspace để sử dụng



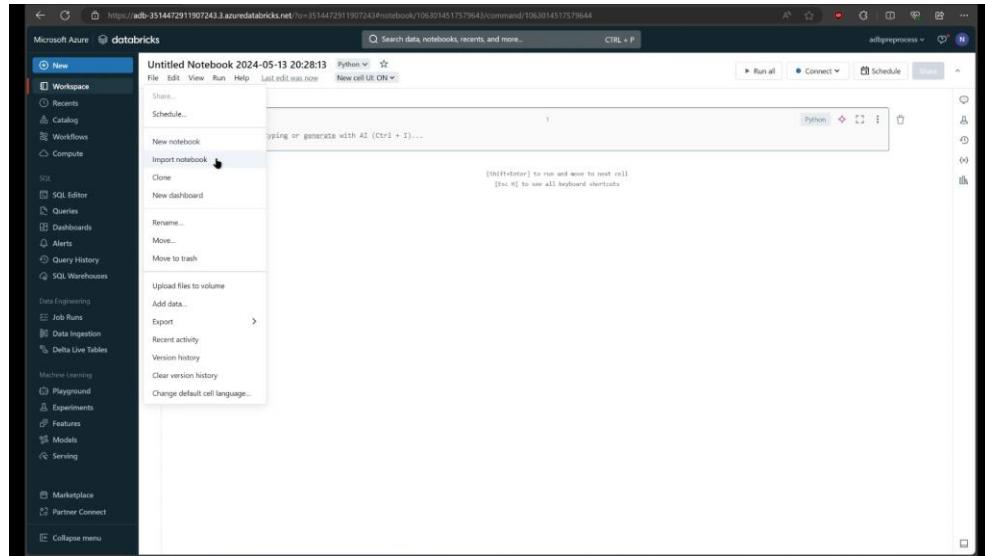
Hình 62. Sau khi đăng nhập, tạo mới Compute bằng cách chọn Create Compute



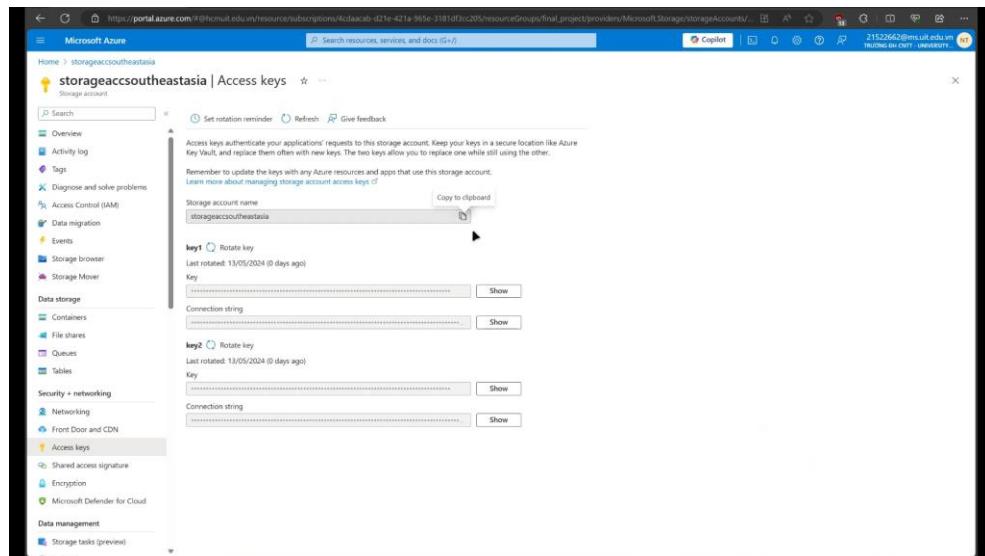
Hình 63. Setup lại Cluster và chọn Create



Hình 64. Tại menu chọn New và Notebook để tạo mới

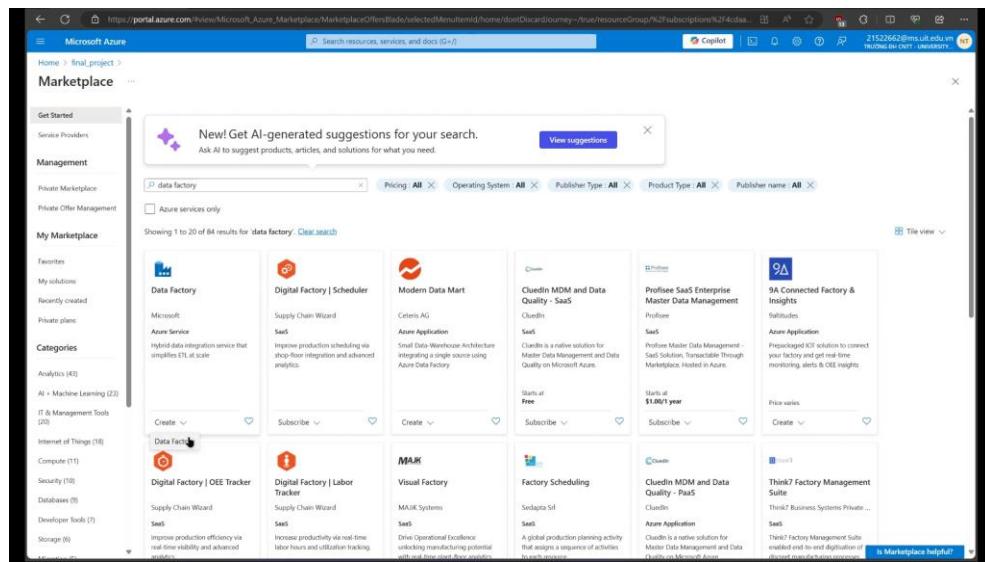


Hình 65. Chọn File và thực hiện Import notebook từ máy

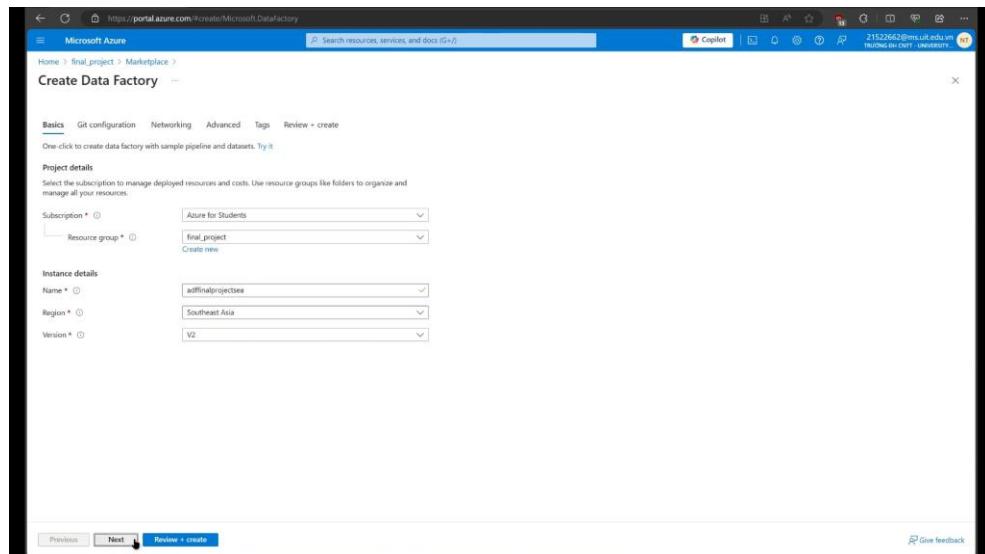


Hình 66. Vào lại Storage Account lấy key và paste vào notebook

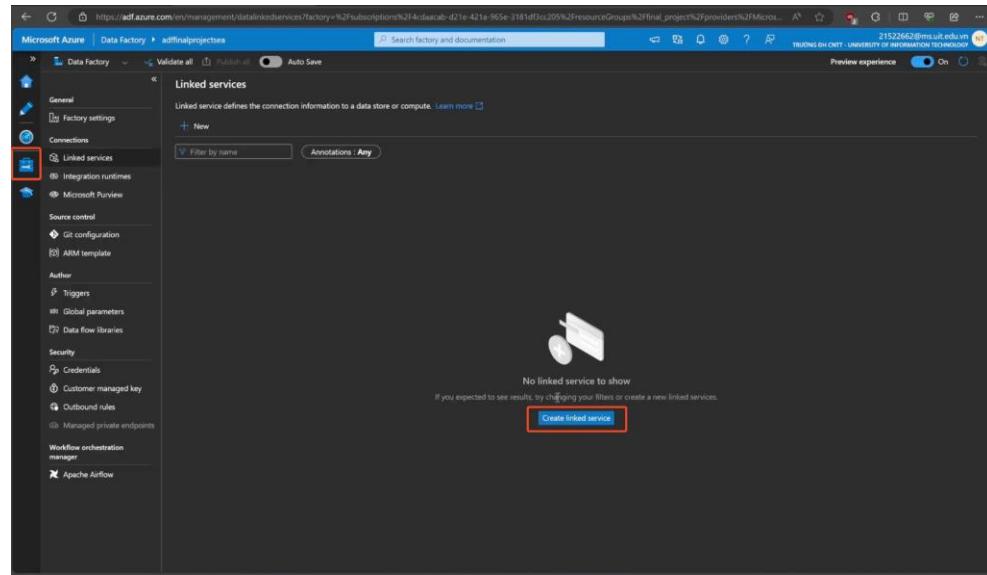
2.1.5. Cài đặt Azure Data Factory



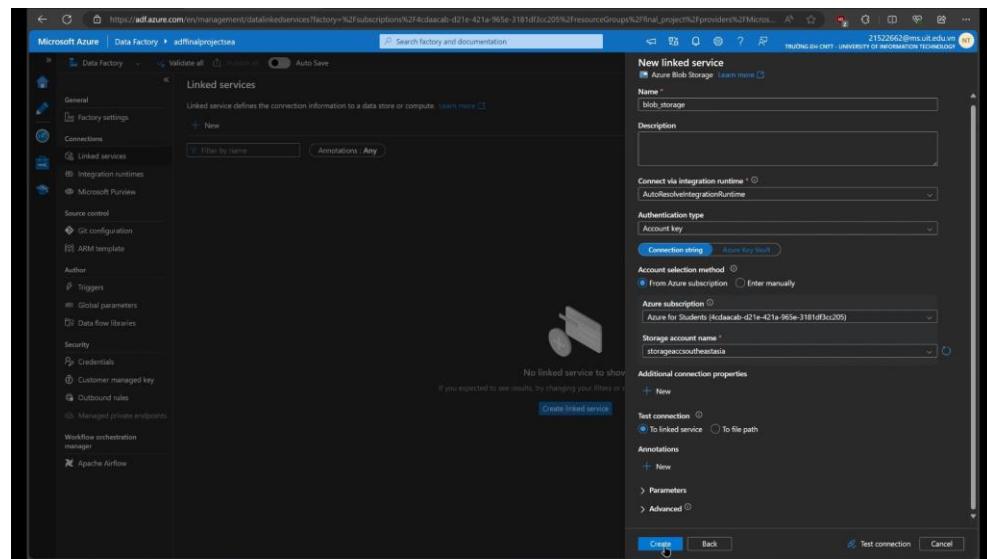
Hình 67. Vào lại Marketplace để tạo mới Data Factory



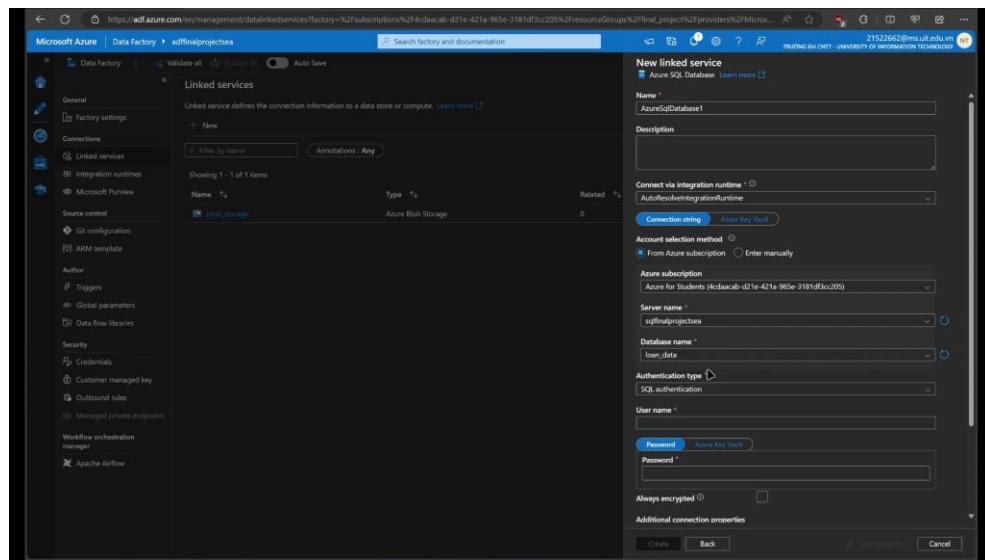
Hình 68. Nhập thông tin các tab và thực hiện Review + Create



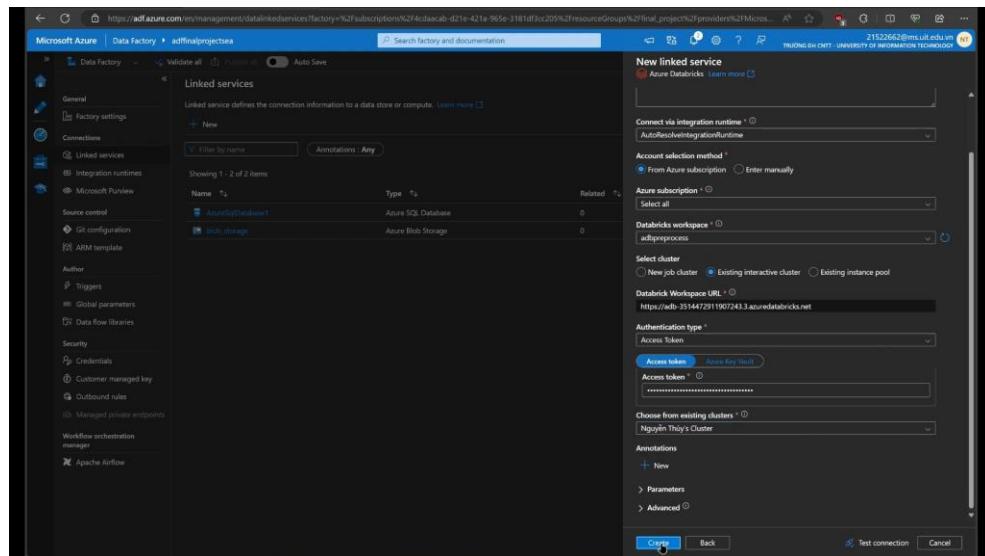
Hình 69. Ở giao diện Data Factory, chuyển tới mục Connections, chọn Linked Services và Create Linked Services để kết nối tới các dịch vụ cần thiết



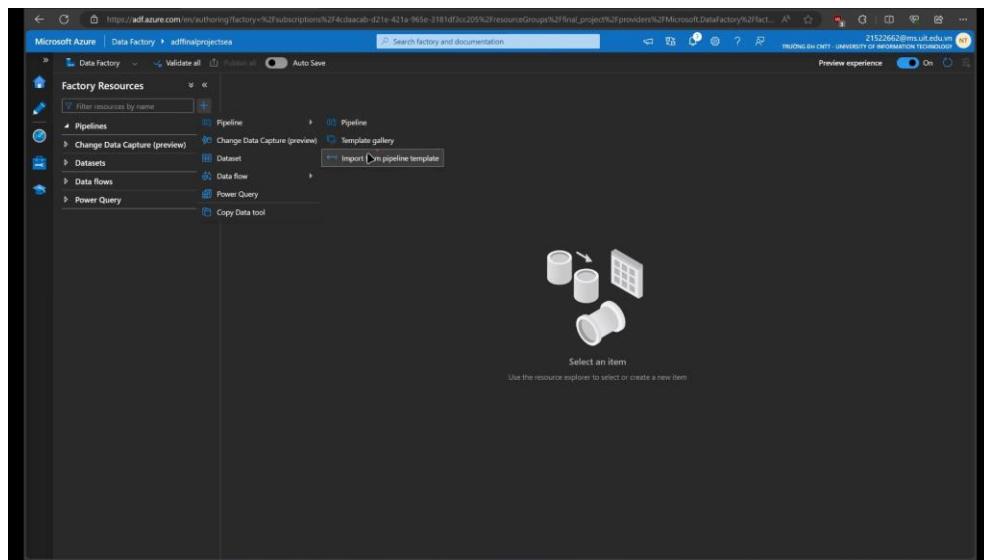
Hình 70. Đầu tiên thiết lập dịch vụ Blob Storage như hình



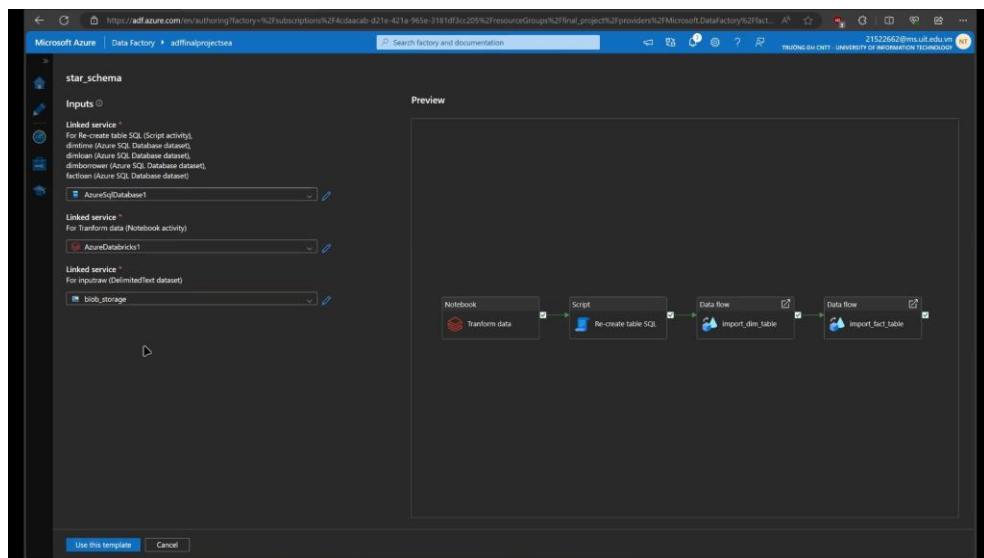
Hình 71. Tiếp tục thiết lập dịch vụ SQL Database và nhập thông tin vào



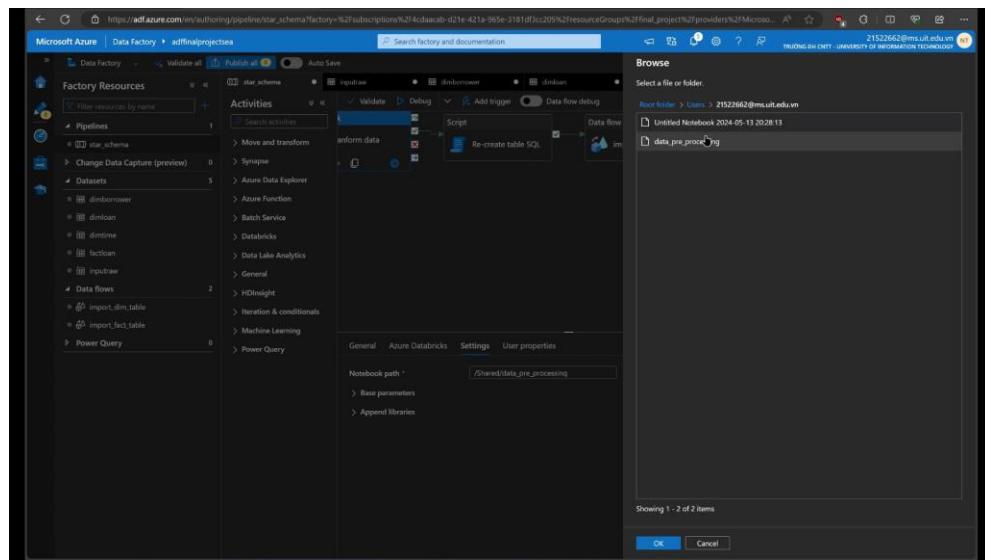
Hình 72. Tiếp tục thiết lập dịch vụ Databricks



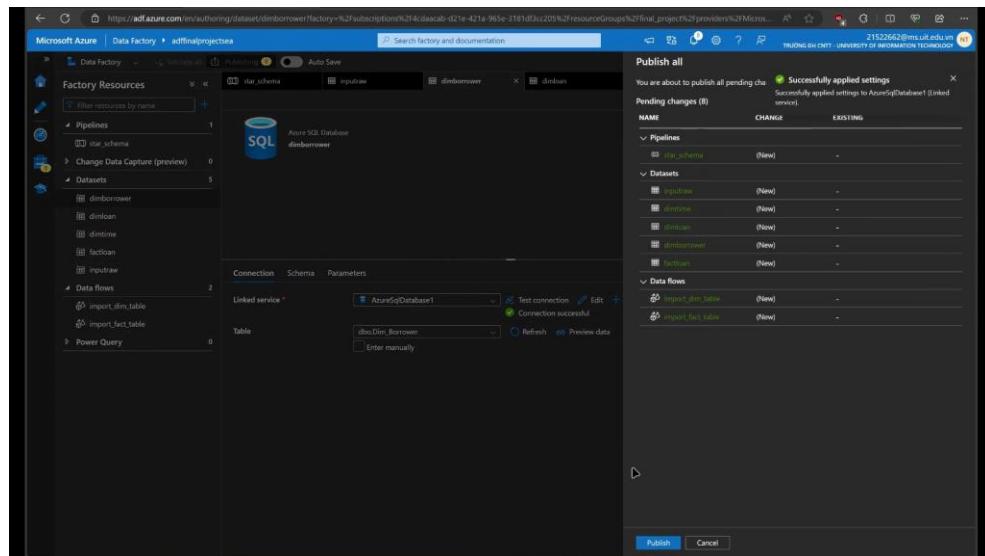
Hình 73. Chọn lại tab Author trong menu và Import from pipeline template từ máy



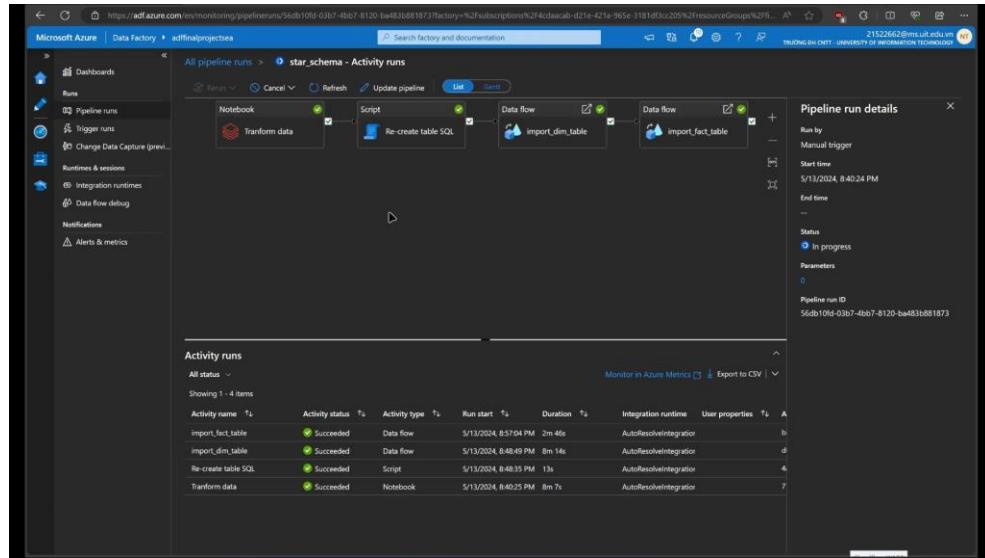
Hình 74. Sau khi import, cấu hình pipeline như trên hình và chọn Use this template



Hình 75. Sau đó, tại activity Transform data, qua tab Settings chọn Notebook path đúng



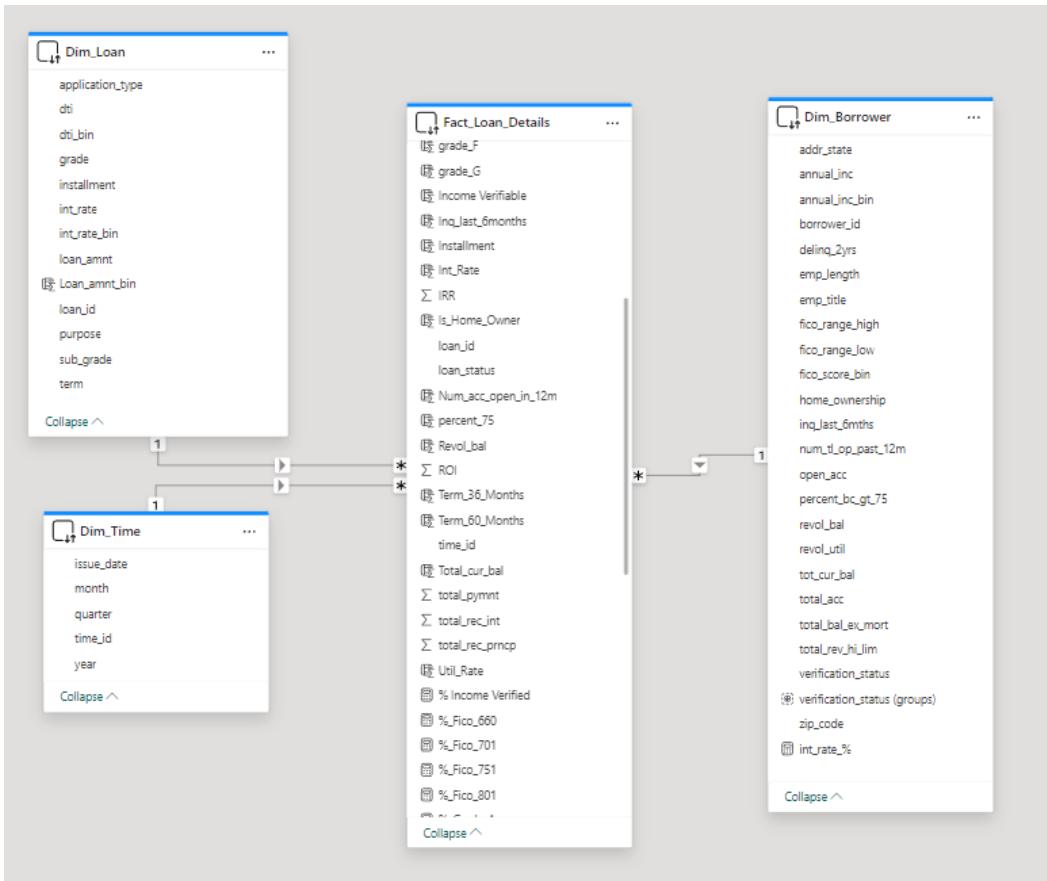
Hình 76. Sau khi kiểm tra qua các activity còn lại, thực hiện Publish all và chọn Publish



Hình 77. Sau khi chọn Add trigger để chạy Pipeline, chờ cho Run thành công

Hình 78. Để kiểm tra xem dữ liệu đã được copy vào bảng trong database, quay lại SQL Database "loan_data", Chọn Query editor (preview) trong phần Settings để chạy thử câu truy vấn và trả kết quả như trên là đã xem như thành công

2.2. Hiện thực trực quan hóa dữ liệu



Hình 79. Data model trong PowerBI sau khi tạo các measure cần thiết cho các report (sử dụng hàm DAX)

The screenshot shows the Role-Based Security (RLS) configuration for the "Long-term Joint Loan Analyst" role:

- Roles** (Left):
 - + New
 - Long-term Joint Lo... (Selected)
 - Manager
 - Mature Loan Analyst
 - Regional Loan Proces...
- Select tables** (Middle):
 - Dim_Borrower
 - Dim_Loan (Selected)
 - Dim_Time
 - Fact_Loan_Det...
- Filter data** (Right):
 - Show data if All of these rules are true
 - Column Condition Value
 - application_type Equals Joint App
 - term Equals 60 months

Hình 80. Cấu hình RLS - Role: Long-term Joint Loan Analyst

The screenshot shows the 'Select tables' and 'Filter data' sections of the Power BI Data Source configuration. In the 'Select tables' section, several tables are listed: Dim_Borrower, Dim_Loan, Dim_Time, Fact_Loan_Det..., Dim_Time, and Fact_Loan_Det... (repeated). In the 'Filter data' section, two rules are defined:

- Rule 1: Show data if All of these rules are true. Condition: grade In [A X, B X, C X].
- Rule 2: Condition: year Is Greater Than 2015.

Hình 81. Cấu hình RLS - Role: Mature Loan Analyst

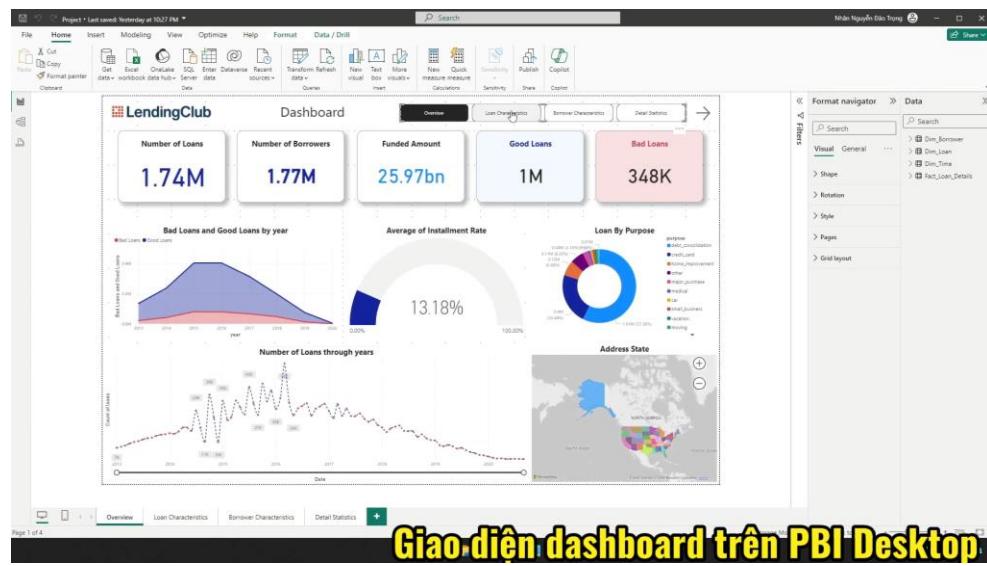
The screenshot shows the 'Select tables' and 'Filter data' sections of the Power BI Data Source configuration. In the 'Select tables' section, several tables are listed: Dim_Borrower, Dim_Loan, Dim_Time, Fact_Loan_Det..., Dim_Time, and Fact_Loan_Det... (repeated). In the 'Filter data' section, one rule is defined:

- Show data if All of these rules are true. Condition: addr_state In [PA X, KY X, IN X, HI X, VA X, KS X, DE X, SC X, DC X, NH X].

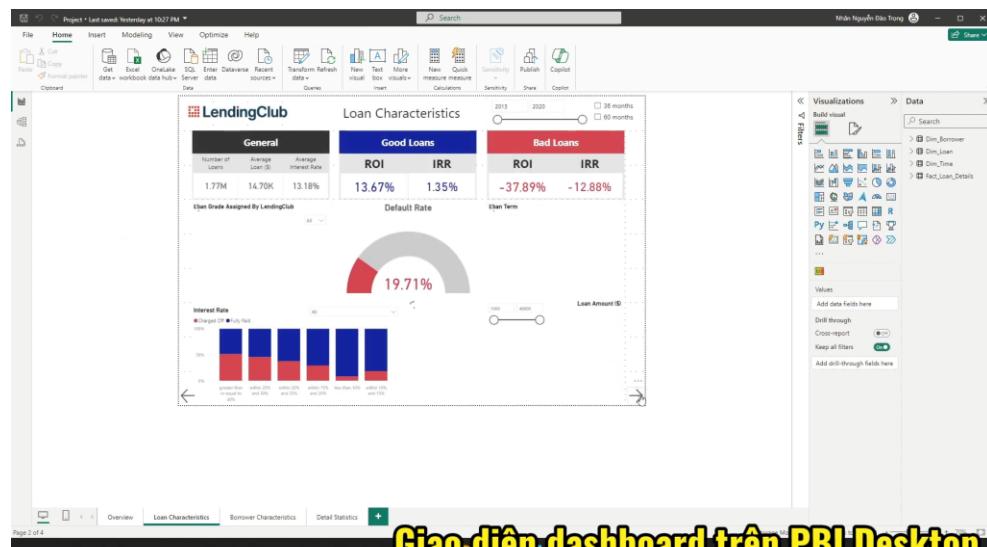
Hình 82. Cấu hình RLS - Role: Regional Loan Processor

The screenshot shows the Power BI desktop interface with a 'Dashboard' view. A 'Build visual' dialog box is open, prompting for a 'SQL Server database' connection. The 'Server' field is set to 'lendingclubdatabase.windows.net'. The 'Data Connectivity mode' dropdown is set to 'DirectQuery'. The 'OK' button is highlighted.

Hình 83. Thực hiện kết nối với Azure SQL DB, nhập đúng tên Server và chọn DirectQuery mode để tạo live connection



Hình 84. Sau khi load các bảng cần thiết thì ta có giao diện dashboard như trên, đầu tiên là report Overview



Hình 85. Report thứ 2: Loan Characteristics



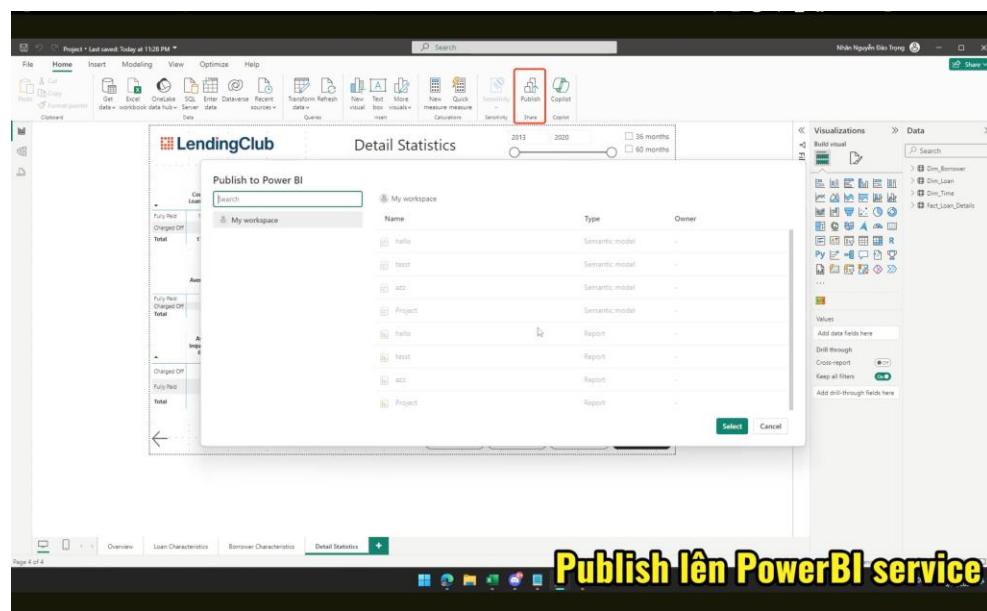
Giao diện dashboard trên PBI Desktop

Hình 86. Report thứ 3: Borrower Characteristics

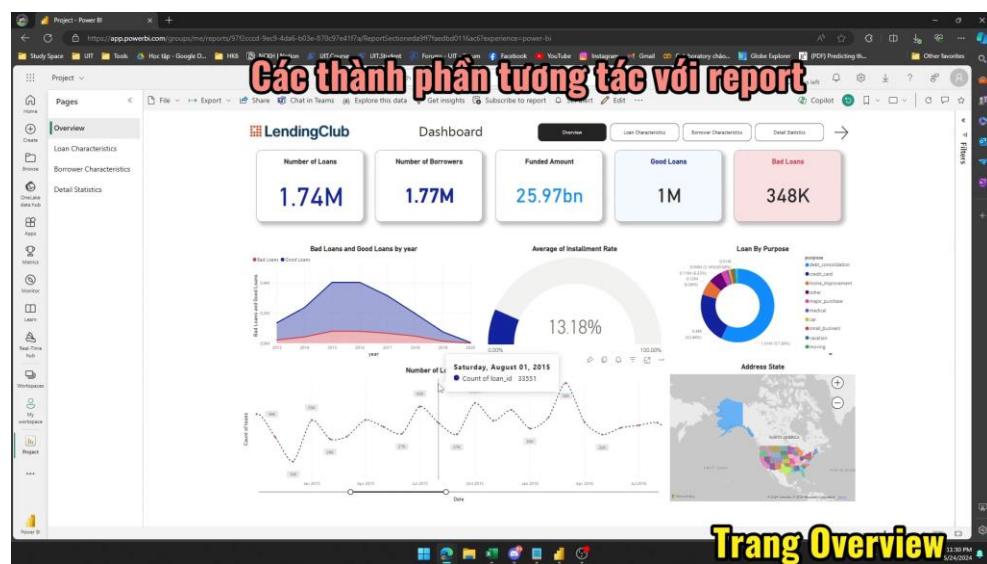


Giao diện dashboard trên PBI Desktop

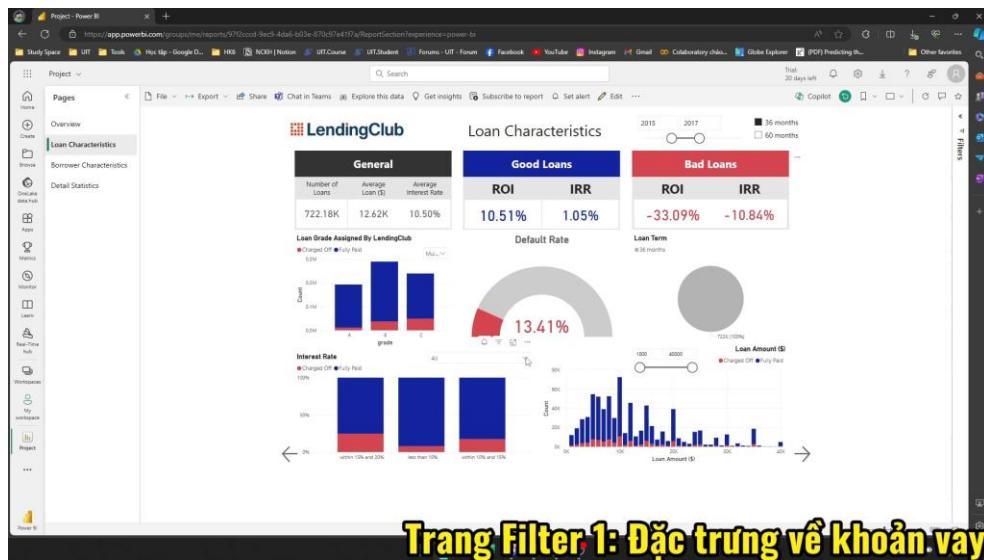
Hình 87. Report 4: Detail Statistics



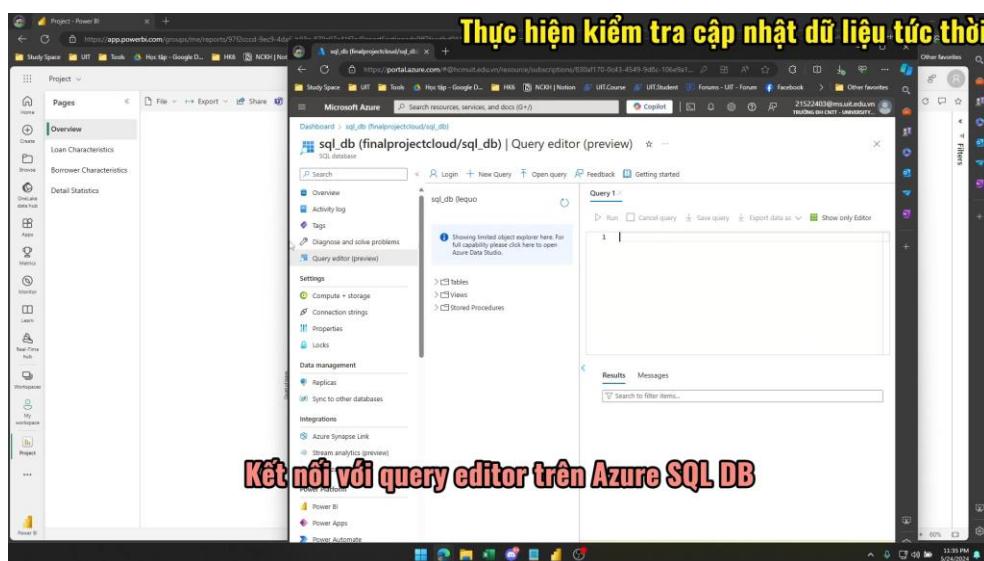
Hình 88. Sau khi save thì thực hiện Publish lên PowerBI service



Hình 89. Kiểm tra report trên PowerBI service đã upload lên chưa và thực hiện tương tác với report



Hình 90. Tương tác thử trên PowerBI service với các report, ví dụ report 2: thực hiện tương tác filter và trả kết quả theo Slicers



Hình 91. Tiếp theo là kiểm tra cập nhật dữ liệu tức thời, đầu tiên vào Azure SQL Database để cập nhật dữ liệu, đăng nhập và chọn Query editor để thực hiện

The screenshot shows the Microsoft Azure Data Studio interface. On the left, the sidebar includes options like Overview, Activity log, Tags, Diagnose and solve problems, Query editor (previews), Settings, Compute + storage, Connection strings, Properties, Logs, Data management, Replicas, Sync to other databases, Integrations, Azure Synapse Link, Stream analytics (preview), Add Azure AI Search, Power Platform, Power BI, Power Apps, and Power Automate. The main area displays two queries:

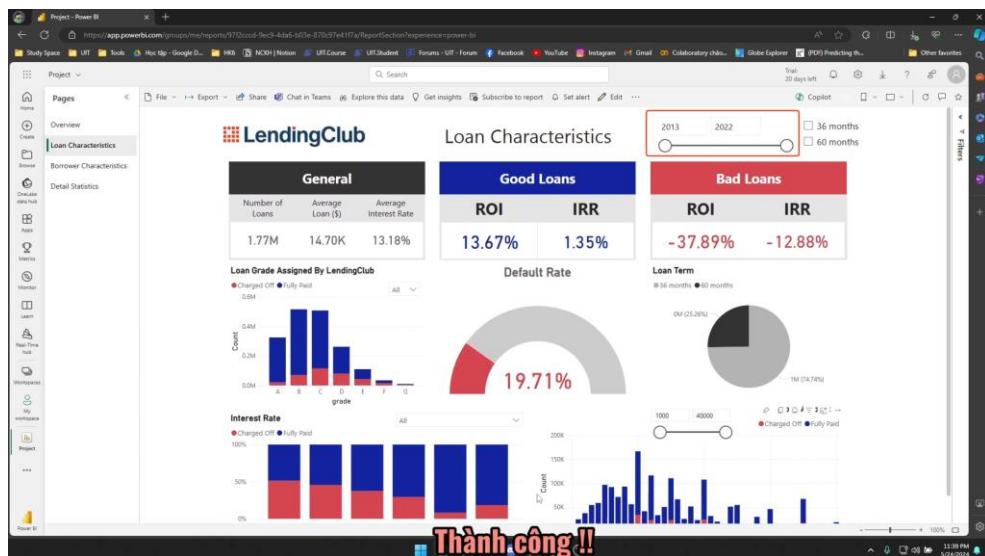
```

Query 1 | Query 2
1. INSERT INTO [dbo].[Dim_Time]([issue_date, month, quarter, year]
2. VALUES
3. ('2021-04-01', 4, 2, 2021),
4. ('2022-05-01', 5, 2, 2022);
5. Select * from [dbo].[Dim_Time]

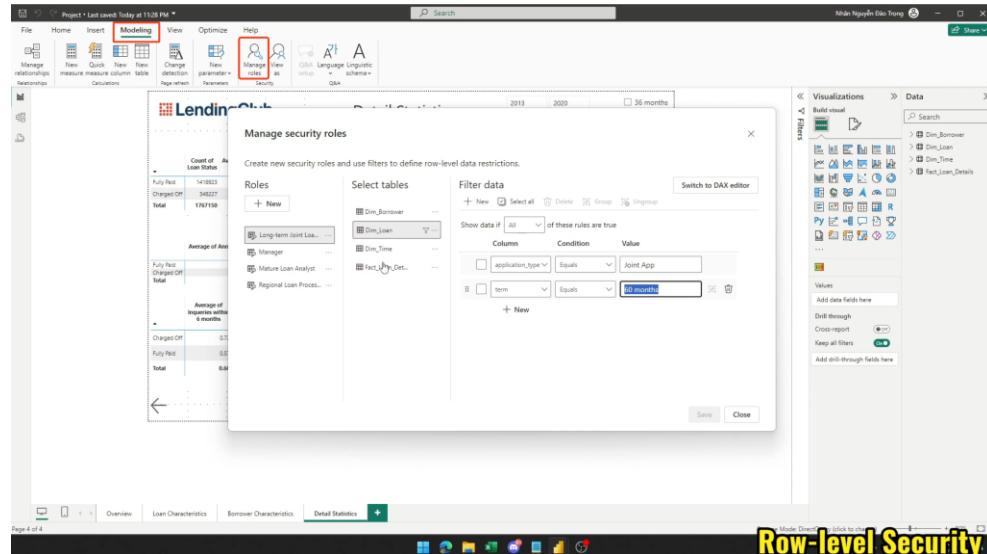
```

The Results tab shows the output of the query, listing dates from 2016-12-01 to 2020-09-01, with columns for issue_date, month, quarter, and year.

Hình 92. Thực hiện truy vấn thêm 2 ngày vào năm 2021 và 2022

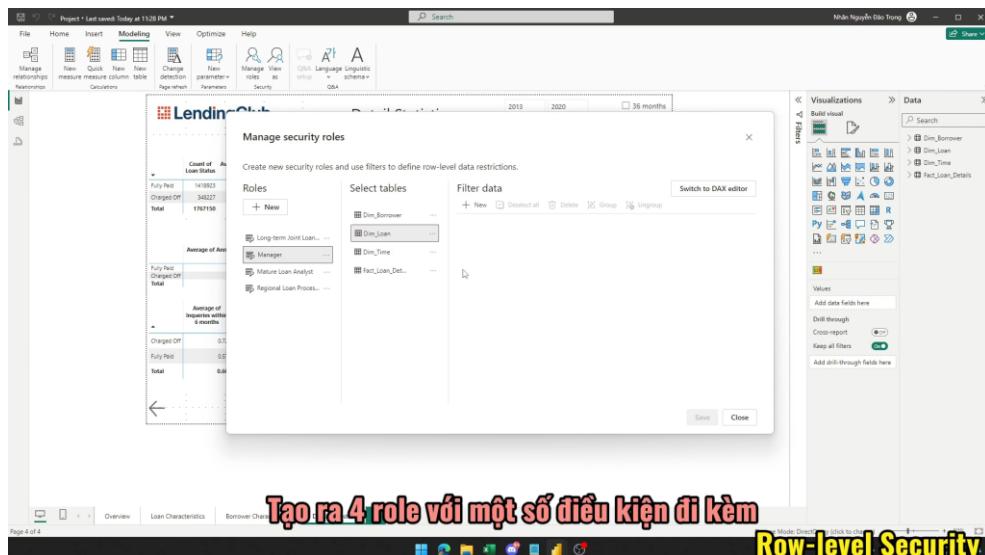


Hình 93. Sau khi Refresh lại trang để dashboard cập nhật, ta có thể thấy thay đổi ở filter đã tăng range từ 2017-2020 lên năm 2017-2022



Row-level Security

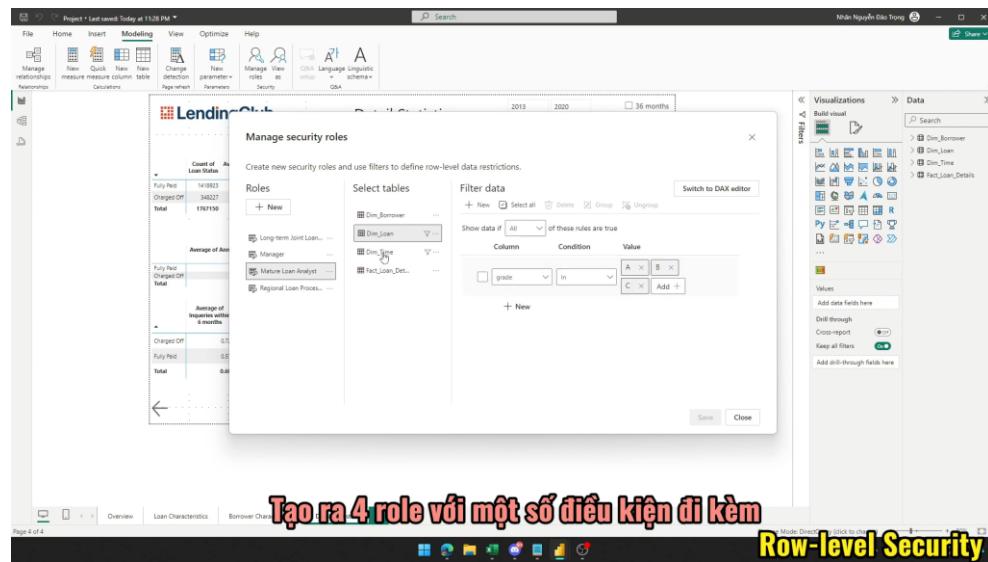
Hình 94. Tiếp theo đến phàn thiết lập quyền bảo mật, chọn vào *Manage roles* trong tab *Modeling* và setup cho role *Long-term Joint Loan Analyst* như hình



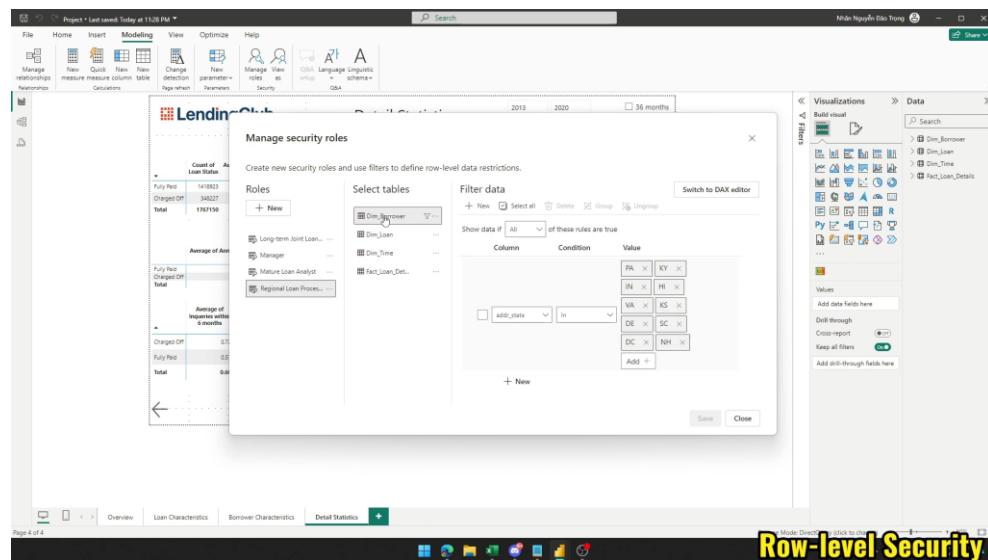
Tạo ra 4 role với một số điều kiện đi kèm

Row-level Security

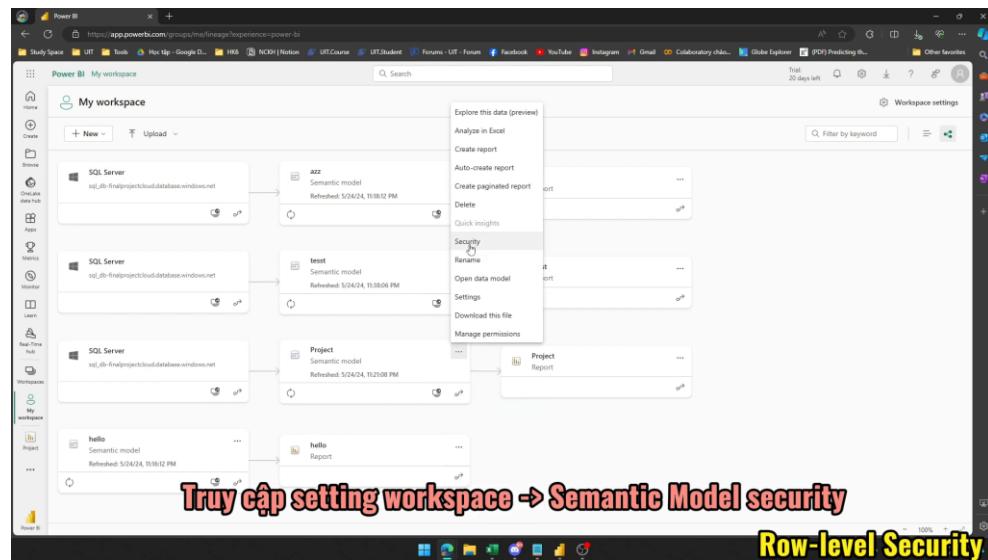
Hình 95. Tạo role Manager với toàn quyền quản lý (không có filter ở các bảng)



Hình 96. Tạo role Mature Loan Analyst thiết lập quyền như hình

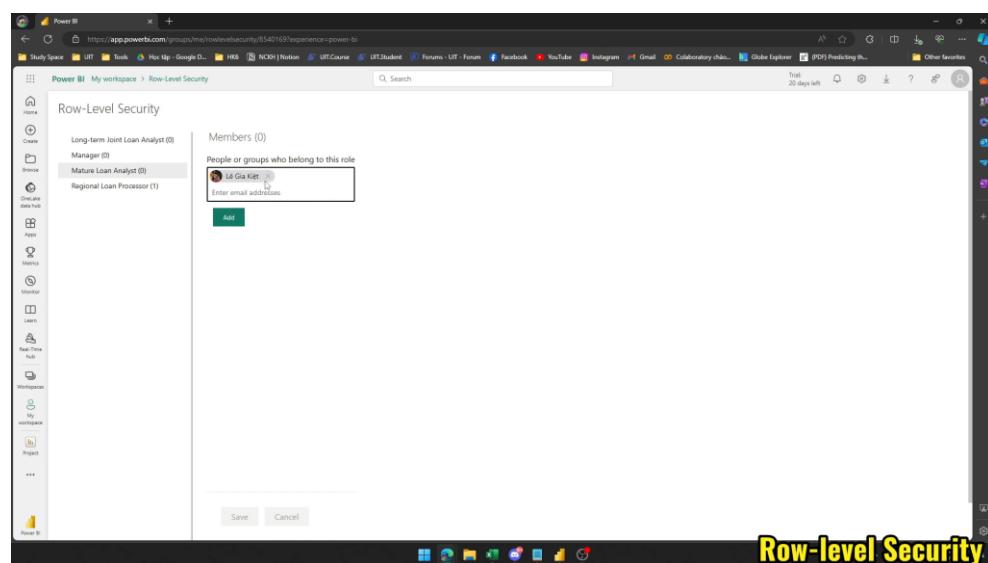


Hình 97. Role Regional Loan Processor có thiết lập quyền như hình

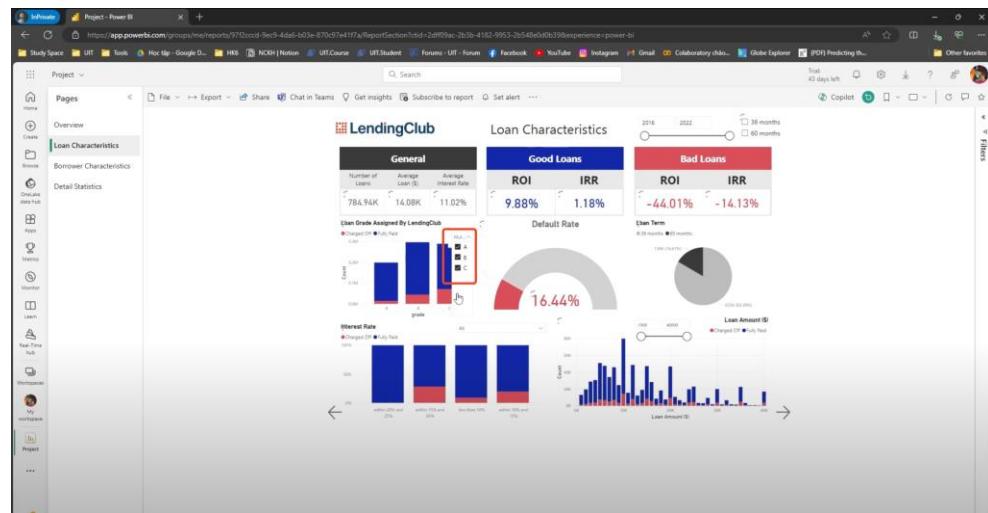


Row-level Security

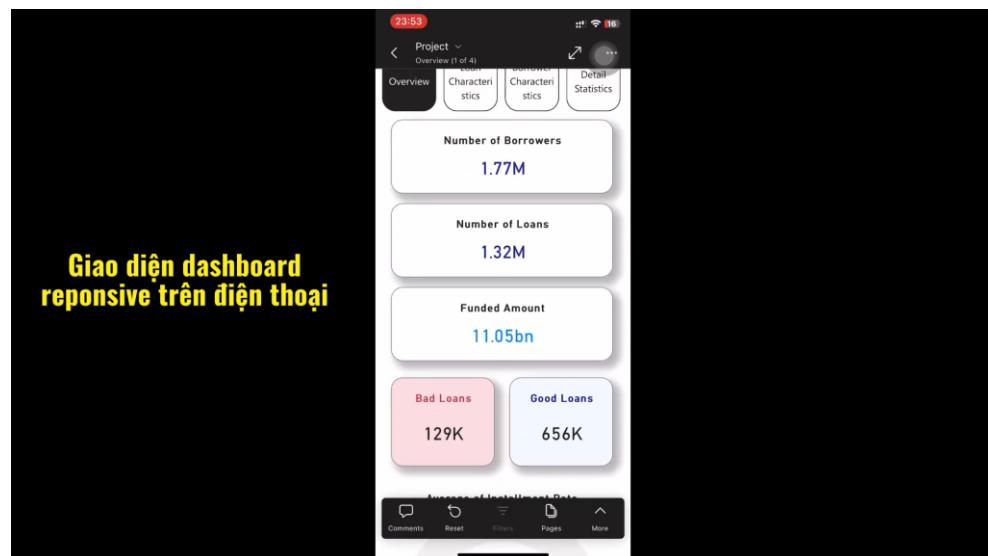
Hình 98. Thực hiện cấp quyền cho user, truy cập vào setting workspace, chọn Security của Semantic Model như hình



Hình 99. Test với role Mature Loan Analyst cấp cho user Lê Gia Kiết và nhấn Add

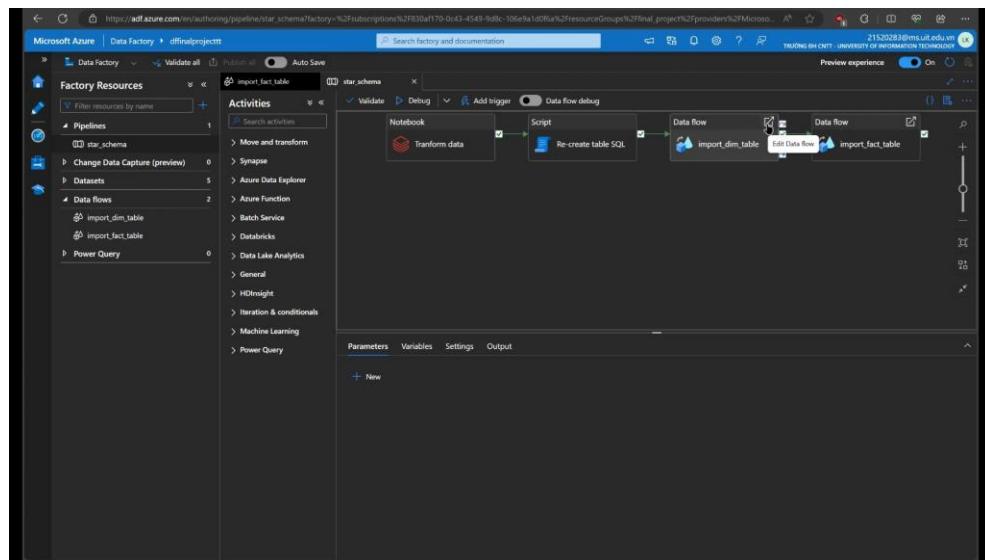


Hình 100. Ta có thể thấy khi đăng nhập user Lê Gia Kiệt đã được cấp role Mature Loan Analyst thì chỉ có thể xem và tương tác trên các giới hạn được cấp (grade là A,B,C)

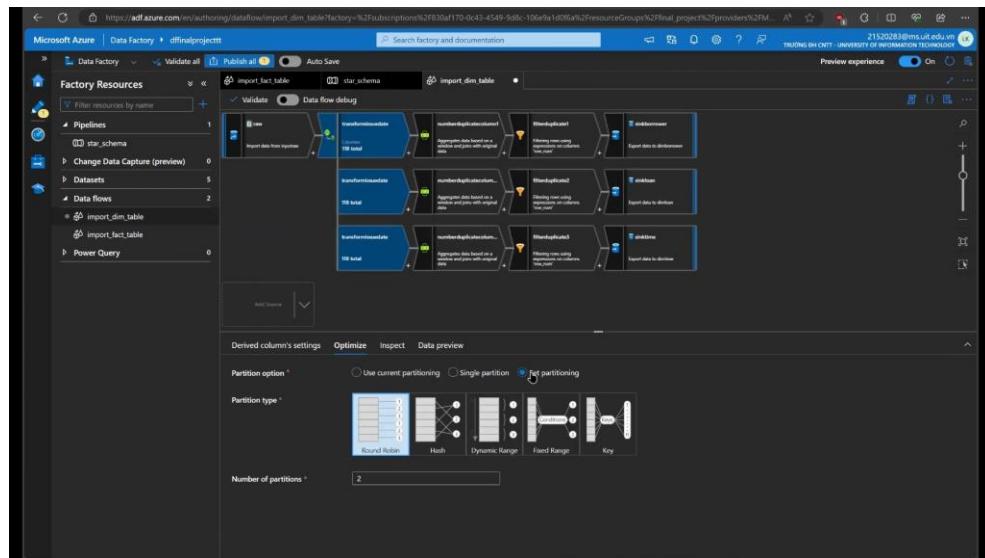


Hình 101. Thực hiện dashboard responsive trên thiết bị điện thoại ở các report, ví dụ Overview

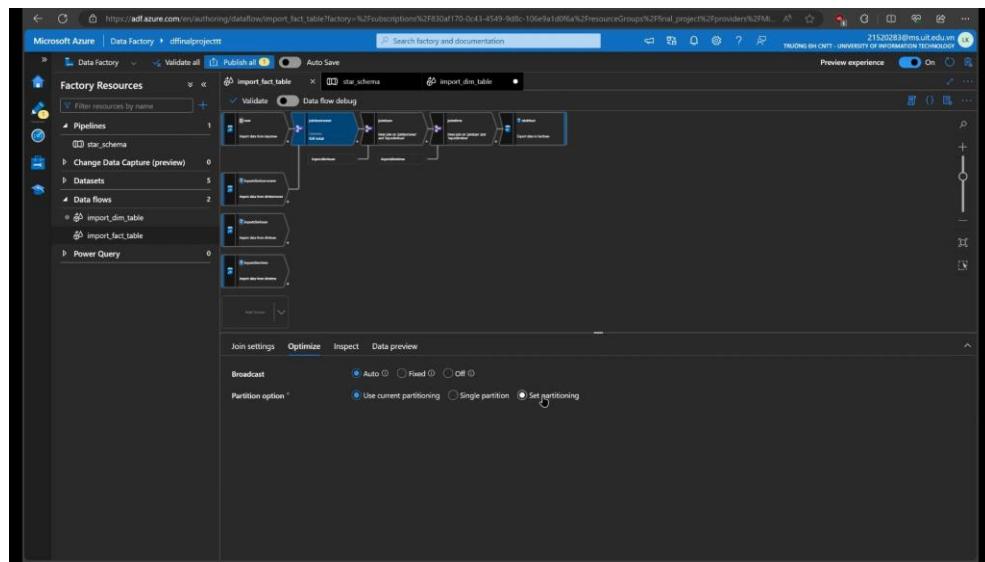
2.3. Giải pháp tối ưu



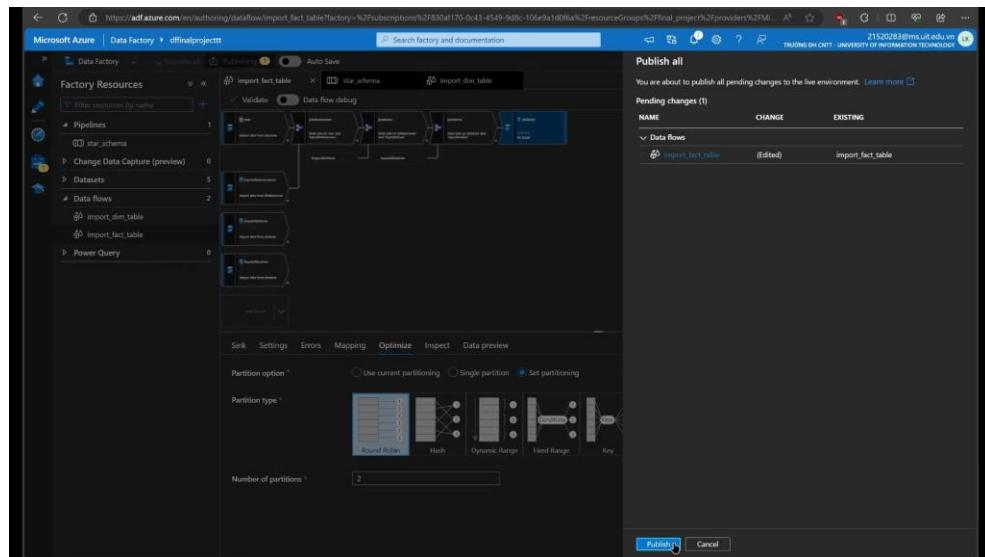
Hình 102. Thực hiện tối ưu trong Data Flows



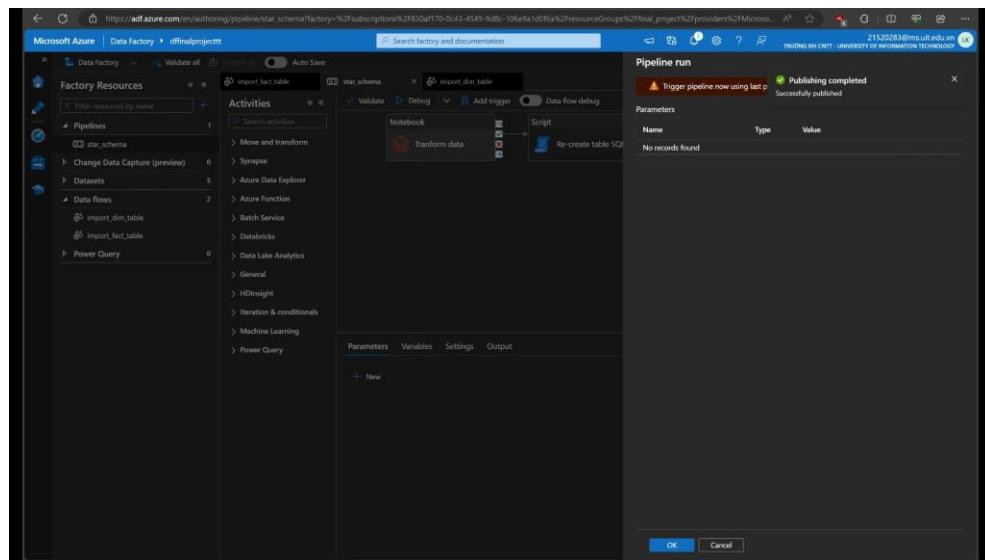
Hình 103. Tối ưu activity transformmississuedate, numberduplicatecolumn1, numberduplicatecolumn2, numberduplicatecolumn3, filterduplicate1, filterduplicate2, filterduplicate3, sinkborrower, sinkloan, sinktime của import_dim_table : Bật phân vùng Set partitioning trong tab Optimize, điều chỉnh sang Round Robin



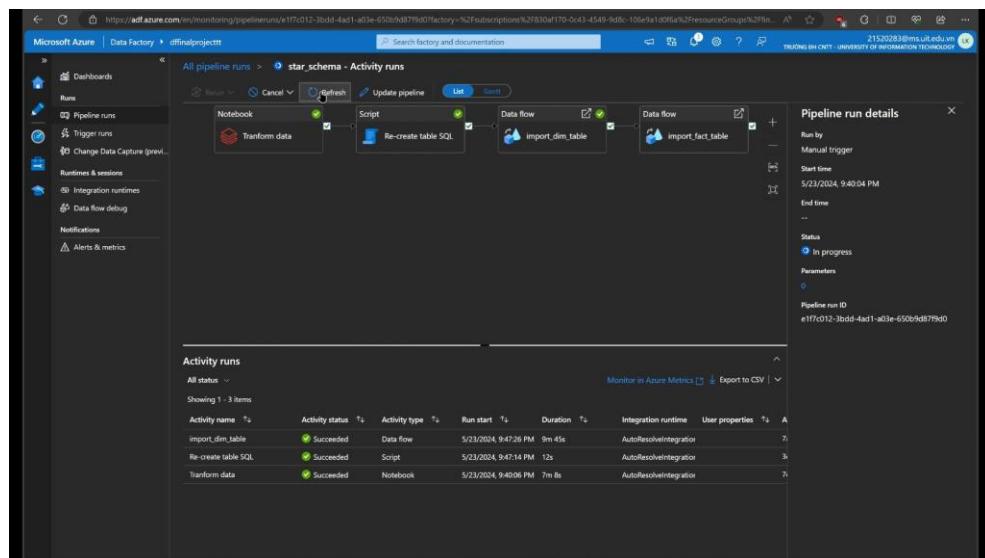
Hình 104. Tối ưu `joinborrower`, `joinloan`, `jointime`, `sinkfact` của `import_fact_table`: Bật phần vùng
Set partitioning trong tab Optimize, điều chỉnh sang Round Robin



Hình 105. Sau đó Publish all



Hình 106. Add trigger để chạy pipeline



Hình 107. Kết quả trả về

THAM KHẢO

- [1] LogicMonitor. (30/5/2023). *What is Azure Blob?* <https://www.logicmonitor.com/blog/what-is-azure-blob>
- [2] Akashdubey-Ms. (10/10/2023). *Introduction to Blob (object) Storage - Azure Storage.* Microsoft Learn. <https://learn.microsoft.com/en-us/azure/storage/blobs/storage-blobs-introduction>
- [3] Team, C. A. (31/1/2023). *Azure SQL: Databases Overview.* Cloud Academy. <https://cloudacademy.com/blog/azure-sql-databases-overview/>
- [4] Senthilkumar, S. (29/9/2023). *Azure Databricks: key features, use cases and benefits.* Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2023/02/azure-databricks-a-comprehensive-guide/>
- [5] Mberdugo. (26/04/2024). *Row-level security (RLS) with Power BI - Microsoft.* Microsoft Learn. <https://learn.microsoft.com/en-us/fabric/security/service-admin-row-level-security>
- [6] Kromerm. (23/10/2023). *Mapping data flow performance and tuning guide - Azure Data Factory & Azure Synapse.* Microsoft Learn. <https://learn.microsoft.com/en-us/azure/data-factory/concepts-data-flow-performance>
- [7] Hoang. (12/04/2024). *DirectQuery in Power BI.* Microsoft Learn. <https://learn.microsoft.com/en-us/power-bi/connect-data/desktop-directquery-about>
- [8] Davidiseminger. (10/11/2023). *DirectQuery in Power BI - Power BI.* Microsoft Learn. <https://learn.microsoft.com/en-us/power-bi/connect-data/desktop-directquery-about>
- [9] Lee, Y. (23/01/2024). *A Journey from LendingClub Data on Kaggle to a Dynamic Loan Analysis Dashboard by Power BI and Python with End-to-End ETL Process.* Medium. <https://medium.com/@yatshunlee/a-journey-from-lendingclub-data-to-a-dynamic-loan-analysis-dashboard-by-power-bi-and-python-with-fb44427ed1f3>

- [10] Popovic, J. (11/05/2023). *Transforming your data in Azure SQL Database to columnstore format*. Microsoft Azure Blog. <https://azure.microsoft.com/en-us/blog/transforming-your-data-in-azure-sql-database-to-columnstore-format/>
- [11] Raunakjhawar. (n.d.). *Online analytical processing (OLAP) - Azure Architecture Center*. Microsoft Learn. <https://learn.microsoft.com/en-us/azure/architecture/data-guide/relational-data/online-analytical-processing>
- [12] Business Intelligence Platform. (2023, November 7). *Choosing the right Azure OLAP service for your business*. Medium. https://medium.com/@Business_Intelligence_Platform/choosing-the-right-azure-olap-service-for-your-business-df2adf455de6
- [13] Jonburchel. (n.d.). *Azure Data Factory Documentation - Azure Data Factory*. Microsoft Learn. <https://learn.microsoft.com/en-us/azure/data-factory/>
- [14] Wikipedia contributors. (2024, February 17). *LendingClub*. Wikipedia. <https://en.wikipedia.org/wiki/LendingClub>