



Bắt đầu với EDQ

Phiên bản 12.2.1

Phòng thí nghiệm thực hành

Phiên bản 1.0

Tác giả:

- Yash Patel (yash.patel@oracle.com)
- David Hecksel (david.hecksel@oracle.com)



Bản quyền © 2015, Oracle và/hoặc các chi nhánh của Oracle. Mọi quyền được bảo lưu.

Oracle ® Enterprise Data Quality, phiên bản 12.2.1

Bản quyền © 2006, 2015, Oracle và/hoặc các chi nhánh của Oracle. Mọi quyền được bảo lưu.

Phần mềm này và tài liệu liên quan được cung cấp theo thỏa thuận cấp phép có chứa các hạn chế về việc sử dụng và tiết lộ và được bảo vệ bởi luật sở hữu trí tuệ. Ngoại trừ khi được cho phép rõ ràng trong thỏa thuận cấp phép của bạn hoặc được pháp luật cho phép, bạn không được sử dụng, sao chép, tái tạo, dịch, phát sóng, sửa đổi, cấp phép, truyền, phân phối, triển lãm, thực hiện, xuất bản hoặc hiển thị bất kỳ phần nào, dưới bất kỳ hình thức nào hoặc bằng bất kỳ phương tiện nào. Việc đảo ngược kỹ thuật, tháo rời hoặc dịch ngược phần mềm này, trừ khi được pháp luật yêu cầu để có khả năng tương tác, đều bị nghiêm cấm.

Thông tin có trong đây có thể thay đổi mà không cần thông báo và không đảm bảo là không có lỗi. Nếu bạn tìm thấy bất kỳ lỗi nào, vui lòng báo cáo cho chúng tôi bằng văn bản.

Nếu đây là phần mềm hoặc tài liệu liên quan được gửi đến Chính phủ Hoa Kỳ hoặc bất kỳ ai cấp phép thay mặt cho Chính phủ Hoa Kỳ, thì thông báo sau đây sẽ được áp dụng:

QUYỀN CỦA CHÍNH PHỦ HOA KỲ Các chương trình, phần mềm, cơ sở dữ liệu và tài liệu liên quan cùng dữ liệu kỹ thuật được cung cấp cho khách hàng của Chính phủ Hoa Kỳ là "phần mềm máy tính thương mại" hoặc "dữ liệu kỹ thuật thương mại" theo Quy định mua sắm liên bang hiện hành và các quy định bổ sung cụ thể của cơ quan. Do đó, việc sử dụng, sao chép, tiết lộ, sửa đổi và điều chỉnh phải tuân theo các hạn chế và điều khoản cấp phép được nêu trong hợp đồng của Chính phủ hiện hành và, trong phạm vi áp dụng theo các điều khoản của hợp đồng của Chính phủ, các quyền bổ sung được nêu trong FAR 52.227-19, Giấy phép phần mềm máy tính thương mại (tháng 12 năm 2007). Oracle America, Inc., 500 Oracle Parkway, Redwood City, CA 94065.

Phần mềm hoặc phần cứng này được phát triển để sử dụng chung trong nhiều ứng dụng quản lý thông tin. Nó không được phát triển hoặc dự định sử dụng trong bất kỳ ứng dụng nào vốn có nguy hiểm, bao gồm các ứng dụng có thể gây ra rủi ro thương tích cá nhân. Nếu bạn sử dụng phần mềm hoặc phần cứng này trong các ứng dụng nguy hiểm, thì bạn phải chịu trách nhiệm thực hiện mọi biện pháp an toàn, sao lưu, dự phòng và các biện pháp khác để đảm bảo sử dụng an toàn. Oracle Corporation và các chi nhánh của Oracle Corporation từ chối mọi trách nhiệm đối với bất kỳ thiệt hại nào do sử dụng phần mềm hoặc phần cứng này trong các ứng dụng nguy hiểm.

Oracle và Java là các nhãn hiệu đã đăng ký của Oracle và/hoặc các chi nhánh của Oracle. Các tên khác có thể là nhãn hiệu của chủ sở hữu tương ứng.

Intel và Intel Xeon là nhãn hiệu hoặc nhãn hiệu đã đăng ký của Intel Corporation. Tất cả các nhãn hiệu SPARC đều được sử dụng theo giấy phép và là nhãn hiệu hoặc nhãn hiệu đã đăng ký của SPARC International, Inc. AMD, Opteron, logo AMD và logo AMD Opteron là nhãn hiệu hoặc nhãn hiệu đã đăng ký của Advanced Micro Devices. UNIX là nhãn hiệu đã đăng ký của The Open Group.

Phần mềm hoặc phần cứng và tài liệu này có thể cung cấp quyền truy cập hoặc thông tin về nội dung, sản phẩm và dịch vụ từ bên thứ ba. Oracle Corporation và các chi nhánh của Oracle không chịu trách nhiệm và từ chối rõ ràng mọi bảo hành dưới bất kỳ hình thức nào liên quan đến nội dung, sản phẩm và dịch vụ của bên thứ ba. Oracle Corporation và các chi nhánh của Oracle sẽ không chịu trách nhiệm cho bất kỳ mất mát, chi phí hoặc thiệt hại nào phát sinh do bạn truy cập hoặc sử dụng nội dung, sản phẩm hoặc dịch vụ của bên thứ ba.

Mục lục

Giới thiệu

Phòng thí nghiệm 1a: Giám đốc Chất lượng dữ liệu doanh nghiệp (EDQ)

Khám phá giao diện người dùng chính xung quanh EDQ, tạo dự án mới và nhập dữ liệu. **Phòng thí nghiệm 1b: Lập hồ sơ dữ liệu của bạn**

Khám phá mối quan hệ dữ liệu, bất thường, sai lệch chuẩn hóa và tác động tiêu cực đến bảng thông tin phân tích kinh doanh quan trọng.

Phòng thí nghiệm 2: Kiểm tra dữ liệu của bạn để xác nhận sự khác biệt. Tạo và sử dụng dữ liệu tham chiếu.

Thực hiện các kiểm tra cụ thể đối với dữ liệu, cách sử dụng Bộ xử lý kiểm toán để thêm cờ và cách chúng cho phép bạn xây dựng các quy trình phân nhánh.

Phòng thí nghiệm 3: Tạo và tự động hóa các giải pháp chuẩn hóa và cải tiến dữ liệu (Kiểm toán, Chuyển đổi, Phân tích)

Chuẩn hóa dữ liệu bằng nhiều bộ xử lý phân tích khác nhau. Thêm hoặc thay thế dữ liệu khi không có để tăng cường tập dữ liệu cho phù hợp để sử dụng.

Phòng thí nghiệm 4: Làm sạch, Phân tích cú pháp, Ghép, Sáp nhập và Loại bỏ trùng lặp

Cải thiện các quy trình được tạo trong các phòng thí nghiệm trước đó để tạo Nguồn thông tin xác thực duy nhất cho các thực thể dữ liệu quan trọng và Phân tích chính xác.

Phòng thí nghiệm 5: Tạo công việc để tự động hóa chất lượng dữ liệu với tích hợp ODI và EDQ

Tạo một công việc để tự động hóa các quy trình được tạo trong các phòng thí nghiệm trước đó. Một ví dụ tiếp theo về tích hợp Oracle Data Integrator và Enterprise Data Quality được đưa vào cuối phòng thí nghiệm này.

Phòng thí nghiệm 6: Quản lý vấn đề

Ghi lại các vấn đề để quản lý chất lượng dữ liệu tốt hơn. Chỉ định các vấn đề cho người dùng khác để tạo không gian làm việc cộng tác nhằm ghi lại và điều tra dữ liệu bất thường. Sử dụng Trình quản lý vấn đề để xem và quản lý các vấn đề.

Mẹo và thủ thuật triển khai

Các biện pháp thực hành tốt nhất để chuẩn bị cho việc triển khai Enterprise Data Quality của bạn. Xem thêm các tài nguyên để hiểu sâu hơn về EDQ.

Giới thiệu

Enterprise Data Quality là giải pháp hàng đầu của Oracle để cung cấp 'Data Fit for Use' và 'Accurate' Analytics đáp ứng các yêu cầu Quản trị dữ liệu hiện tại và mới nổi của tổ chức cho các môi trường/sáng kiến Data Warehouse/Data Mart/Data Lake của bạn. Oracle Enterprise Data Quality cung cấp cho các tổ chức một bộ công cụ chất lượng dữ liệu tích hợp, cung cấp giải pháp toàn diện để đo lường, cải thiện và quản lý chất lượng dữ liệu từ bất kỳ miền nào, bao gồm Khách hàng/Công dân/Sinh viên/Nhân viên/Bệnh nhân và các miền khác. Oracle Enterprise Data Quality cũng kết hợp các khả năng lập hồ sơ, làm sạch, đối sánh và giám sát dữ liệu mạnh mẽ đồng thời mang lại sự dễ sử dụng vô song. Các tính năng của Oracle Enterprise Data Quality bao gồm:

- Phân tích dữ liệu nâng cao để xác định và đo lường dữ liệu chất lượng kém và xác định các yêu cầu về quy tắc để giải quyết các vấn đề về chất lượng dữ liệu của Dự án và/hoặc Doanh nghiệp của bạn
- Nhận dạng dựa trên ngữ nghĩa và mẫu để phân tích chính xác và chuẩn hóa dữ liệu có cấu trúc kém
- Kiến trúc dữ liệu tham chiếu mở sáng tạo cho phép dễ dàng tạo, tùy chỉnh và duy trì các quy tắc kinh doanh thích ứng và học hỏi từ dữ liệu của bạn để cho phép và đẩy nhanh các giải pháp tự động nhằm khởi động, cải thiện liên tục và xã hội hóa chất lượng dữ liệu của bạn theo thời gian
- Tích hợp "Nút dễ dàng" với Oracle Data Integrator Enterprise Edition

Lợi ích chính của việc áp dụng Chất lượng dữ liệu doanh nghiệp

- Tường lửa chất lượng dữ liệu: Cung cấp các quy trình tự động để phát hiện, sửa chữa và/hoặc thiết lập quy trình "trả lại cho người gửi" trên dữ liệu đến từ các hệ thống hoặc đối tác khác trước khi đưa vào Kho dữ liệu quan trọng và/hoặc các tài sản DataMart của bạn
- Hiệu quả về chi phí: Loại bỏ việc "sửa lỗi" dữ liệu thủ công tốn kém và các giải pháp tạm thời khác của nhân viên DBA/Chủ doanh nghiệp
- Báo cáo và Tuân thủ: Cung cấp một nền tảng nhất quán và hợp tác để quản lý chất lượng dữ liệu (bao gồm số liệu DQ Firewall và DQ theo thời gian) bằng cách thực thi các tiêu chuẩn dữ liệu và cho phép tất cả những người tham gia dự án trên khắp các hệ thống và quy trình cùng hiểu các quy tắc DQ

Oracle Enterprise Data Quality Profiling cho tích hợp dữ liệu

Oracle Enterprise Data Quality Profiling for Data Integration cung cấp cơ sở để hiểu các vấn đề về chất lượng dữ liệu và nền tảng để xây dựng các quy tắc chất lượng dữ liệu nhằm khắc phục và ngăn ngừa lỗi. EDQ cho phép người dùng hiểu dữ liệu của họ bằng cách khám phá, làm nổi bật và truyền đạt các bất thường về dữ liệu trong dữ liệu đang được lập hồ sơ. Tạo, truyền đạt, điều tra, cộng tác và đóng các Sự cố về chất lượng dữ liệu. EDQ Data Profiling cung cấp Tường lửa chất lượng dữ liệu cho Kho dữ liệu, Mart hoặc các tài sản dữ liệu quan trọng khác của bạn. Tại sao lại chỉ đưa 'bất kỳ dữ liệu nào' vào Kho dữ liệu và Mart quan trọng đối với doanh nghiệp của bạn? Các ví dụ bao gồm: Sinh viên đại học trong độ tuổi từ -47 đến 114; sinh viên hiện đang theo học tại hệ thống của trường đại học nhưng đã đóng cửa vào năm trước; 27% bệnh nhân trong vòng 10 năm qua sinh ngày 1 tháng 1 năm 1970 một cách bí ẩn. Tất cả những dữ liệu đó đều hợp lệ theo quan điểm của cơ sở dữ liệu - dữ liệu được chèn và cập nhật hoàn toàn bình thường - nhưng liệu đó có thực sự là 'Dữ liệu phù hợp để sử dụng' không? Liệu Analytics có thực hiện được điều đó không

dữ liệu có phải là 'Phân tích chính xác' không? Người dùng lập hồ sơ có thể dễ dàng cộng tác với những người khác trong tổ chức bằng cách mở và chỉ định Sự cố cho Chủ doanh nghiệp, DBA, Nhà khoa học dữ liệu và những người khác trong giao diện người dùng EDQ trong khi khám phá và phát hiện ra các điểm bất thường.

Oracle Enterprise Data Quality Audit và Dashboard để tích hợp dữ liệu

Oracle Enterprise Data Quality Audit and Dashboard for Data Integration cung cấp:

- Nhiều bộ xử lý Kiểm toán linh hoạt và có thể mở rộng để xác thực định dạng, nội dung và sự tuân thủ của dữ liệu trong tay đối với các quy tắc kinh doanh. Các bản ghi có nội dung cột không tuân thủ có thể dễ dàng được tách ra để xử lý / sửa lỗi và/hoặc thông báo đường dẫn lỗi riêng biệt. Bộ xử lý Kiểm toán chỉ được thêm vào quy trình làm việc Chất lượng dữ liệu và được cấu hình trực quan - không cần mã hóa.
- Bảng thông tin và mẫu bảng thông tin được xây dựng sẵn để báo cáo về số liệu DQ trong bảng thông tin điều hành dễ hiểu và có thể định cấu hình được xem trong trình duyệt web yêu thích của bạn. Các nhóm DQ, giám đốc điều hành, nhà cung cấp dữ liệu nguồn, nhóm ETL, thành viên nhóm Trung tâm quản trị dữ liệu xuất sắc và người tiêu dùng của 'Dữ liệu phù hợp để sử dụng' có thể xem cả bảng thông tin so sánh thời điểm và chu kỳ lăn. Cuối cùng, bảng thông tin bổ sung cho giao tiếp của nhóm quy trình công việc Sự cố khi nhóm DQ mở rộng, nhà tài trợ và các bên quan tâm có thể xem lại kết quả Bảng thông tin mới nhất được công bố khi các dự án triển qua các giai đoạn Phát triển, Đảm bảo chất lượng và SDLC tiền sản xuất.

Xử lý hàng loạt Oracle Enterprise Data Quality để tích hợp dữ liệu

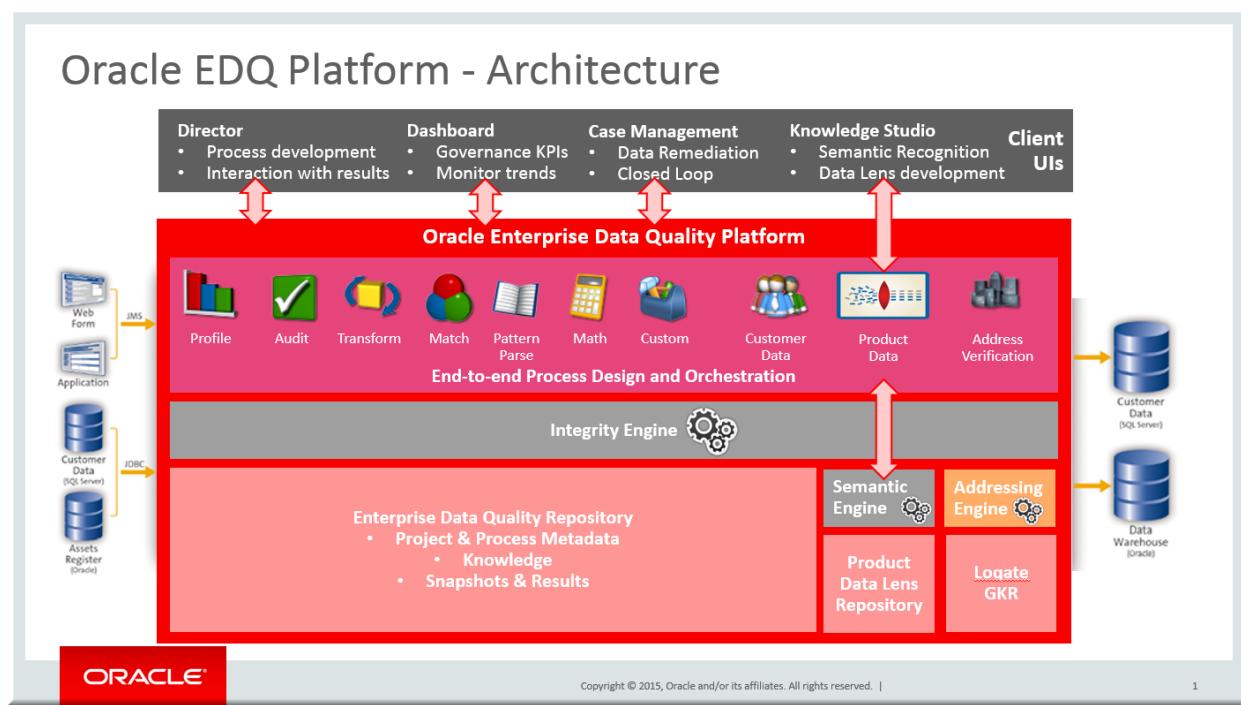
Oracle Enterprise Data Quality Batch Processing for Data Integration cung cấp:

- Nhiều bộ xử lý chuyển đổi linh hoạt và có thể mở rộng để chuẩn hóa và làm giàu dữ liệu nguồn thành dữ liệu tuân thủ các quy tắc kinh doanh và yêu cầu của dự án. Ví dụ, một bộ dữ liệu Khách hàng có 18 cách biểu diễn/cách viết khác nhau của thành phố 'Toronto'. Toronto, toronto, TORONTO, TOR, tor, torontto, ... Tận dụng bộ Dữ liệu tham chiếu mở rộng của EDQ và Kiến trúc dữ liệu tham chiếu mở 'bí mật' của EDQ, Khách hàng đã sử dụng một bộ xử lý Tra cứu và Trả về duy nhất và trả nó đến tập dữ liệu tham chiếu Toronto và tất cả các cách biểu diễn không hợp lệ đều được chuyển đổi thành 'Toronto' – đơn giản hóa công việc của nhà phát triển Trí tuệ kinh doanh và tăng thêm tính toàn vẹn của dữ liệu.
- Một bảng màu phong phú và có thể cấu hình của Bộ xử lý trực quan để thêm xử lý bản ghi khớp, hợp nhất và loại bỏ trùng lặp vào quy trình làm việc DQ

Bây giờ dữ liệu đã được xác thực, chuyển đổi, làm giàu và chuẩn hóa, có thể thực hiện xử lý khớp/hợp nhất/loại bỏ trùng lặp tùy chọn. Lab 4 tập trung vào việc thêm chức năng khớp, hợp nhất và loại bỏ trùng lặp vào quy trình làm việc của Hands-on Lab.

Kiến trúc thành phần chất lượng dữ liệu doanh nghiệp Oracle

Oracle Enterprise Data Quality (OEDQ) là một Ứng dụng Web Java có lợi thế là có một ứng dụng khách phong phú để tạo quy trình công việc DQ Process, tạo/xem Bảng thông tin, lập hồ sơ dữ liệu, kiểm tra, phân tích nâng cao, chuẩn hóa, thực hiện xử lý phê duyệt Đánh giá phù hợp và Quản trị sản phẩm (cùng nhiều tính năng khác).



Phòng thí nghiệm 1a: Giám đốc chất lượng dữ liệu doanh nghiệp

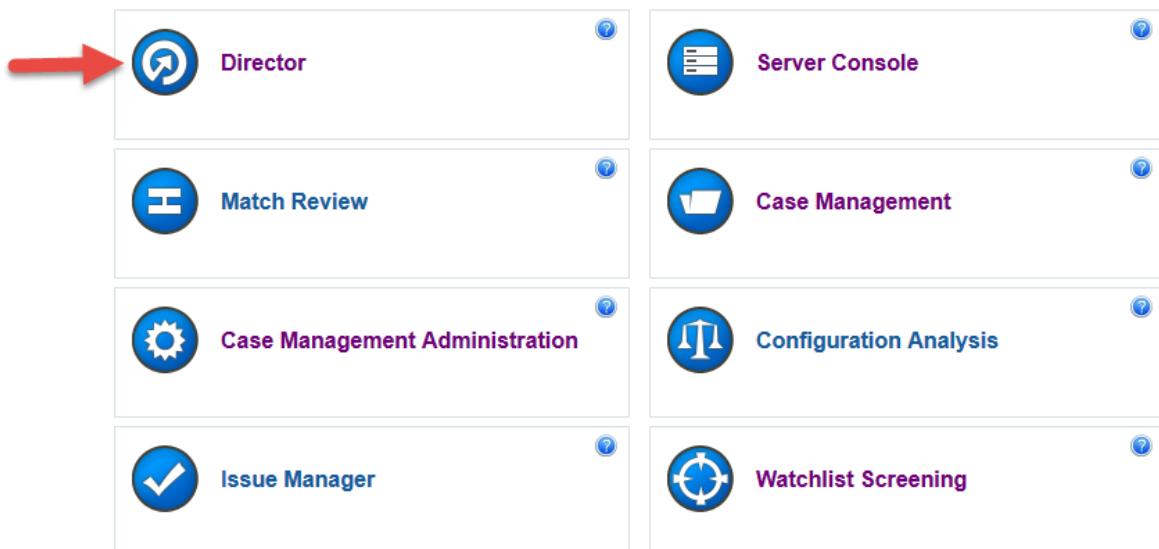
Điều hướng đến Enterprise Data Quality Launchpad và bắt đầu Director

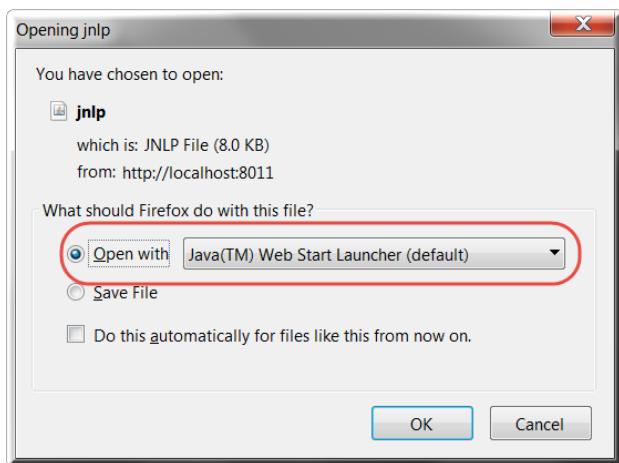
- Khởi chạy trình duyệt internet (Mozilla Firefox, Internet Explorer, Google Chrome) bằng hệ điều hành máy chủ của bạn và điều hướng đến URL <http://localhost:8011/edq>



Lưu ý rằng hệ điều hành máy chủ của bạn sẽ yêu cầu cài đặt môi trường chạy Java 7 mới nhất (JRE 1.7.0_55+). Ngoài ra, bạn có thể hoàn thành hội thảo này bằng cách mở trình duyệt Mozilla Firefox trong máy ảo và điều hướng đến một URL hơi khác: <http://localhost:8001/edq>

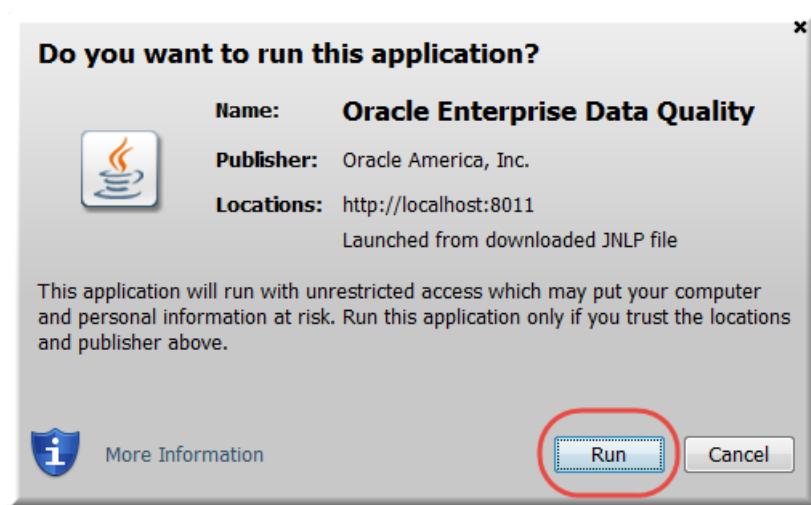
- Nhấp vào biểu tượng Director để khởi chạy giao diện người dùng Director





3. Bạn có thể được nhắc mở một tệp. Chọn **mở** và chọn **Bắt đầu Web Java ứng dụng** và nhấp chuột **ĐƯỢC RỒI**

i *Có thể mất vài phút để tải xuống ứng dụng web. Bạn có thể được nhắc cho phép ứng dụng chạy, hãy nhấp vào **Chạy***

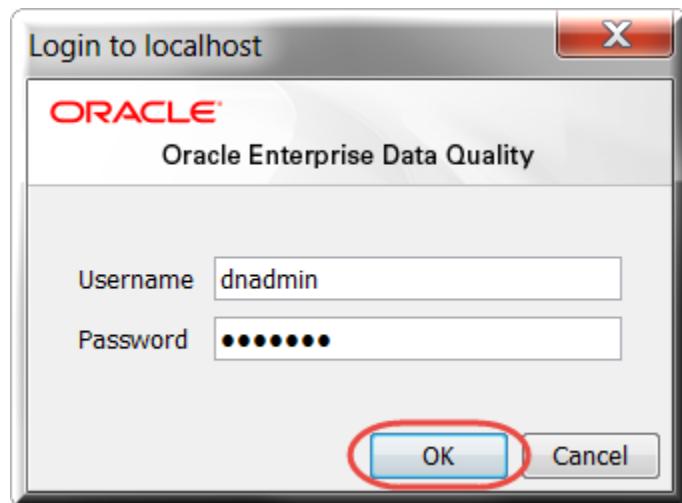


4. Bạn sẽ được nhắc đăng nhập vào Enterprise Data Quality. Sử dụng thông tin đăng nhập sau:

Tên người dùng: dnadmin

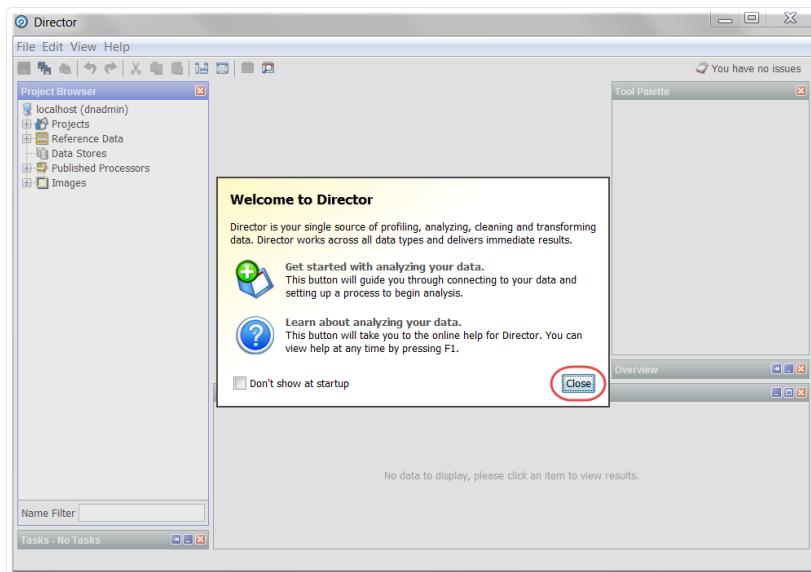
Mật khẩu: dnadmin

Sau đó nhấp vào **ĐƯỢC RỒI**

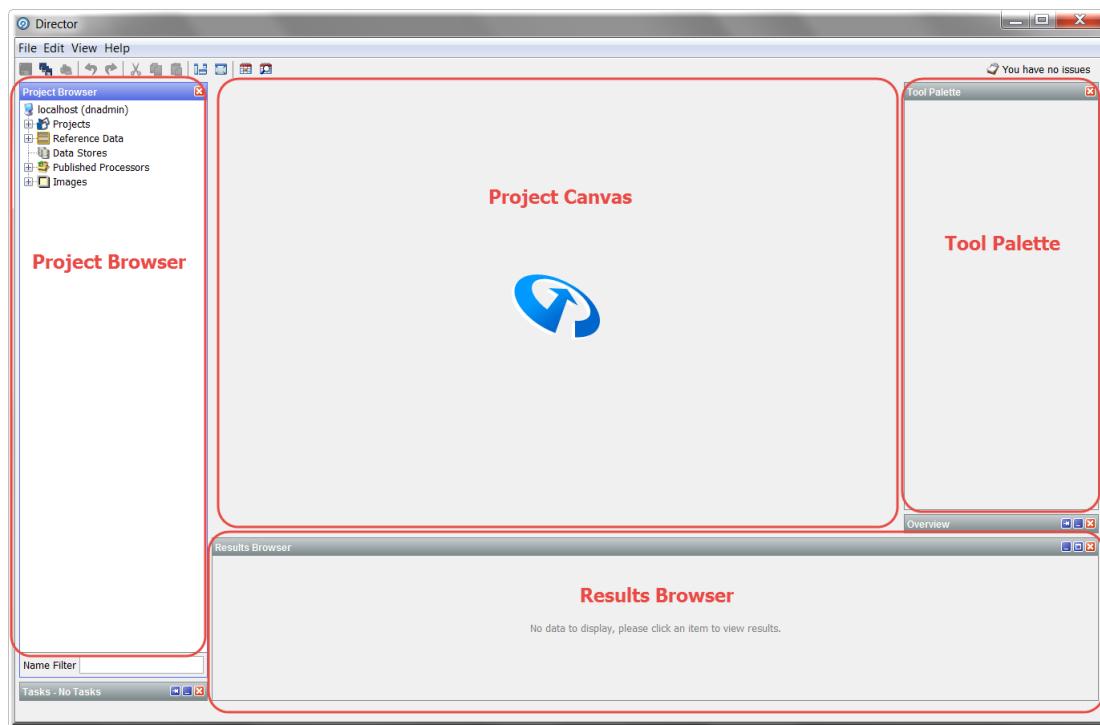


Khám phá Giao diện người dùng của Giám đốc

5. Sau khi Director khởi chạy, hãy nhấp vào **đóng** trong **Chào mừng đến với Giám đốc** tin nhắn:



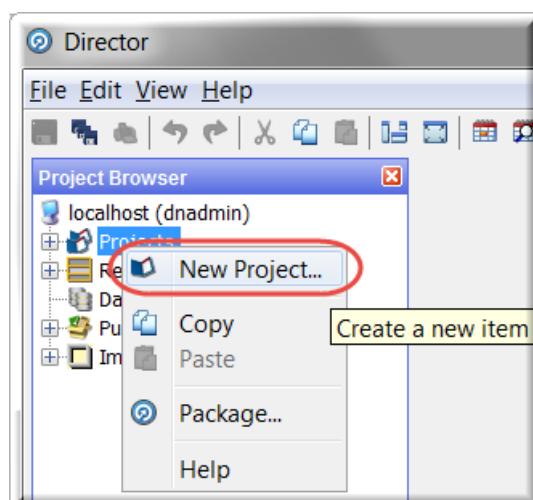
6. Hãy dành chút thời gian để làm quen với Thuật ngữ của từng lĩnh vực khác nhau trong bốn lĩnh vực **Giám đốc** ứng dụng



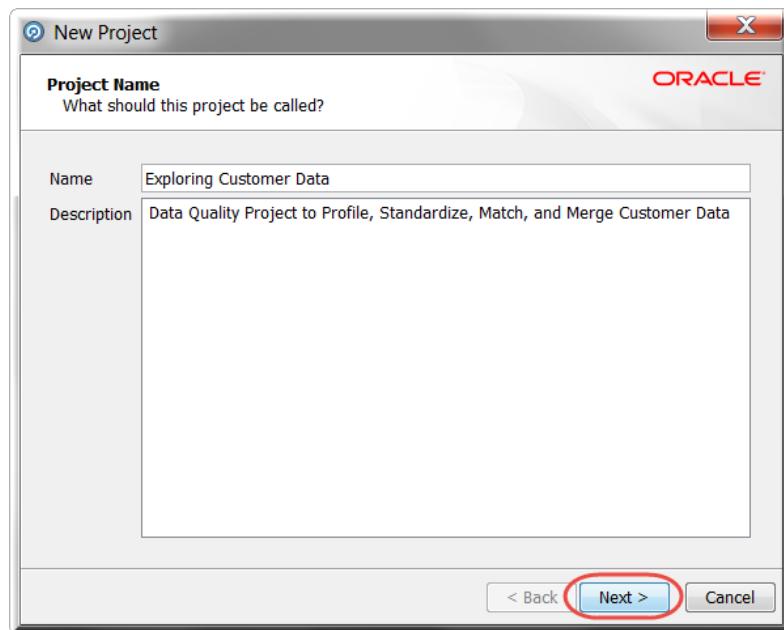
Tạo một dự án mới

Các dự án được tạo trong Project Brower và thường được sử dụng để lưu trữ dữ liệu và quy trình liên quan đến sáng kiến Data Quality. Bạn có thể thiết lập quyền và mức truy cập ở cấp độ dự án. Bây giờ chúng ta sẽ bắt đầu một dự án chất lượng dữ liệu sử dụng dữ liệu khách hàng mẫu từ hệ thống quản lý đơn hàng có trụ sở tại Hoa Kỳ trong suốt phần còn lại của phòng thí nghiệm.

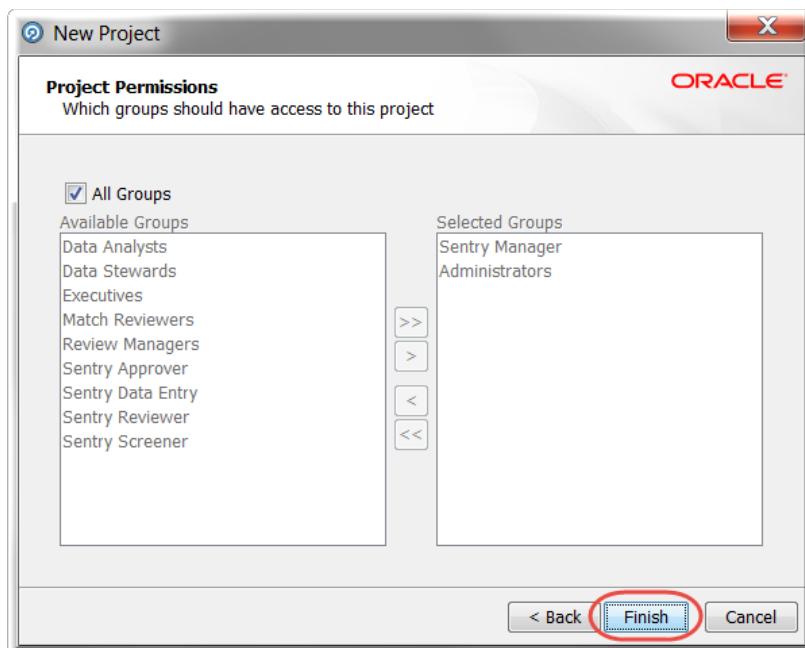
- Trong **Trình duyệt dự án**, nhấp chuột phải **Dự án** và chọn **Dự án mới...** để bắt đầu Wizard



2. Đối với **Tên**, đi vào *Khám phá dữ liệu khách hàng* và tùy chọn thêm một **Sự miêu tả**. Nhấp chuột **Tiếp theo >** để tiếp tục



3. Đảm bảo **Tất cả các nhóm** hợp kiểm được chọn trong **Quyền của dự án**. Điều này sẽ đảm bảo bất kỳ người dùng nào cũng có thể xem và sử dụng dự án. Nhấp vào **Hoàn thành** để tạo dự án mới



Các **Khám phá dữ liệu khách hàng** Dự án hiện xuất hiện trong **Dự án danh sách!**

Thêm kho dữ liệu

Bây giờ chúng ta đã tạo một Dự án cho Phòng thí nghiệm, bước tiếp theo là cho phép truy cập vào Dữ liệu của bạn cần Hồ sơ hóa/Làm giàu/Chuẩn hóa và tùy chọn Khớp/Gộp/Loại bỏ trùng lặp. Biến Dữ liệu của bạn thành:

- 'không chỉ là Dữ liệu', mà là 'Dữ liệu Phù hợp để Sử dụng'
- 'không chỉ là Phân tích', mà là 'Phân tích Chính xác'

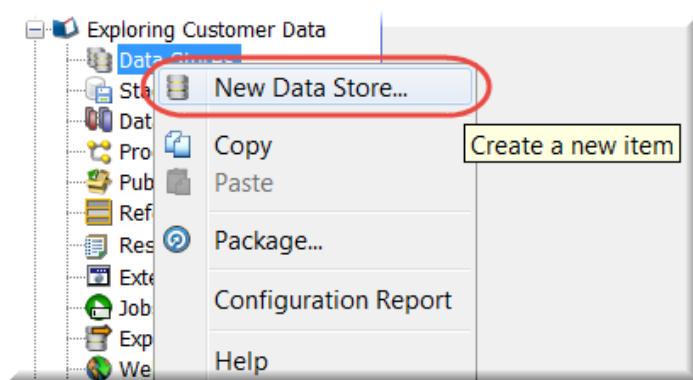
Kho dữ liệu là kết nối đến kho dữ liệu, cho dù dữ liệu được lưu trữ trong cơ sở dữ liệu hay trong một hoặc nhiều tệp. Kho dữ liệu có thể được sử dụng làm nguồn dữ liệu cho một quy trình hoặc bạn có thể xuất kết quả Dữ liệu theo giai đoạn đã viết của một quy trình sang kho dữ liệu hoặc cả hai.



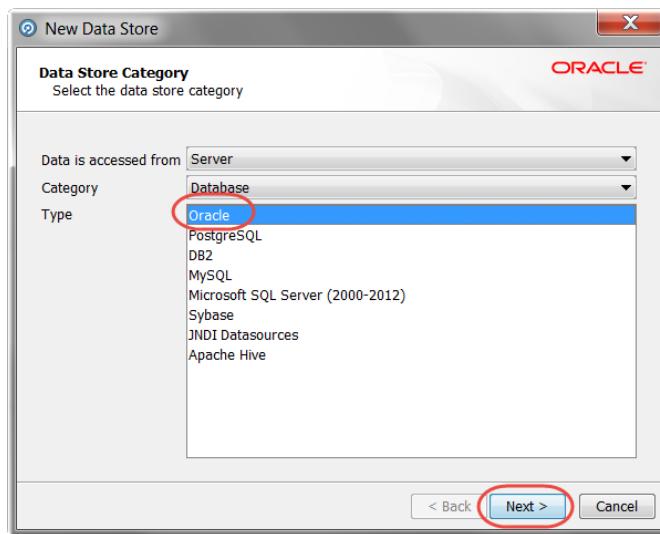
Thông thường, nên kết nối với kho dữ liệu thông qua máy chủ. Khi kết nối với tệp, điều này có nghĩa là tệp phải tồn tại trong vùng đích của máy chủ để đảm bảo máy chủ có thể truy cập chúng. Tuy nhiên, cũng có thể kéo dữ liệu vào máy chủ bằng kết nối máy khách.

EDQ hỗ trợ kết nối gốc đến nhiều loại kho dữ liệu. Hãy bắt đầu bằng cách thêm kho dữ liệu để kết nối với Cơ sở dữ liệu Oracle:

1. Mở rộng dự án mới tạo **Khám phá dữ liệu khách hàng**, nhấp chuột phải **Kho dữ liệu** và chọn **Kho dữ liệu mới** để khởi chạy **Kho dữ liệu mới** Phù thủy



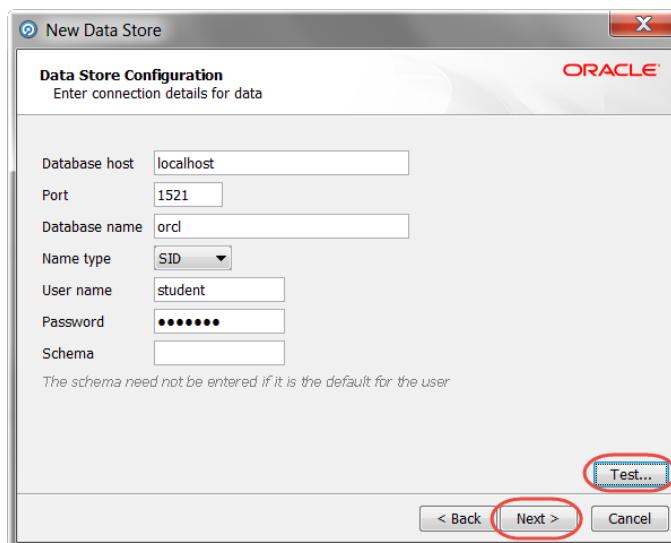
2. Chọn **Máy chủ** và **Cơ sở dữ liệu** trong hai hộp thả xuống, sau đó chọn **Tiếp theo >** để tiếp tục



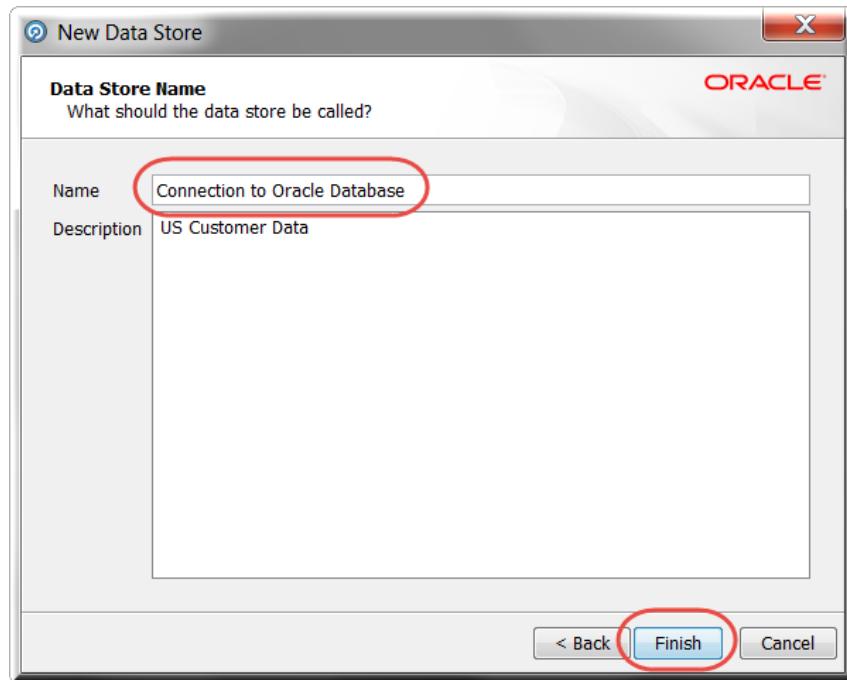
3. Nhập các thông tin kết nối sau:

- **Máy chủ cơ sở dữ liệu:** *máy chủ cục bộ*
- **bộ Cảng:** *1521*
- **Tên cơ sở dữ liệu**(Mã số thuế:*orcl*)
- **Tên người dùng:***học sinh* **Mật khẩu:***học sinh*

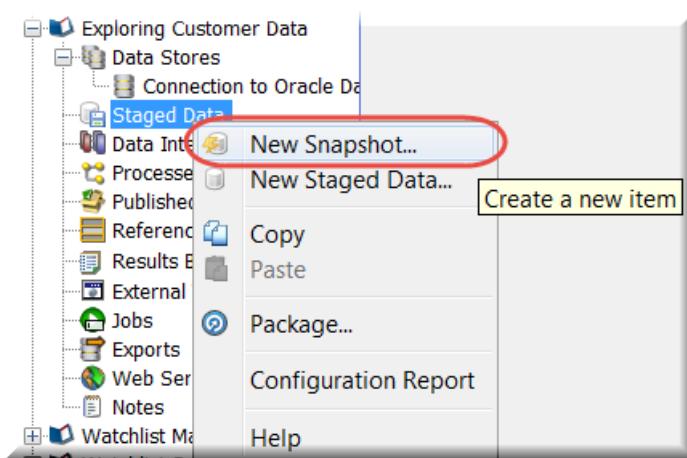
Nhấp chuột **Bài kiểm tra** để kiểm tra xem Director có thể truy cập vào Cơ sở dữ liệu máy chủ hay không, sau đó nhấp vào **Tiếp theo >** để tiếp tục



4. Bây giờ là lúc đặt tên cho Kho dữ liệu của chúng ta. Đối với **Tên** trường nhập văn bản, nhập Kết nối với cơ sở dữ liệu Oracle. Bạn có thể tùy chọn nhập văn bản vào vùng văn bản Mô tả để hỗ trợ giao tiếp và cộng tác trên nguồn dữ liệu dự định trong Mô tả, nhấp vào **Hoàn thành**

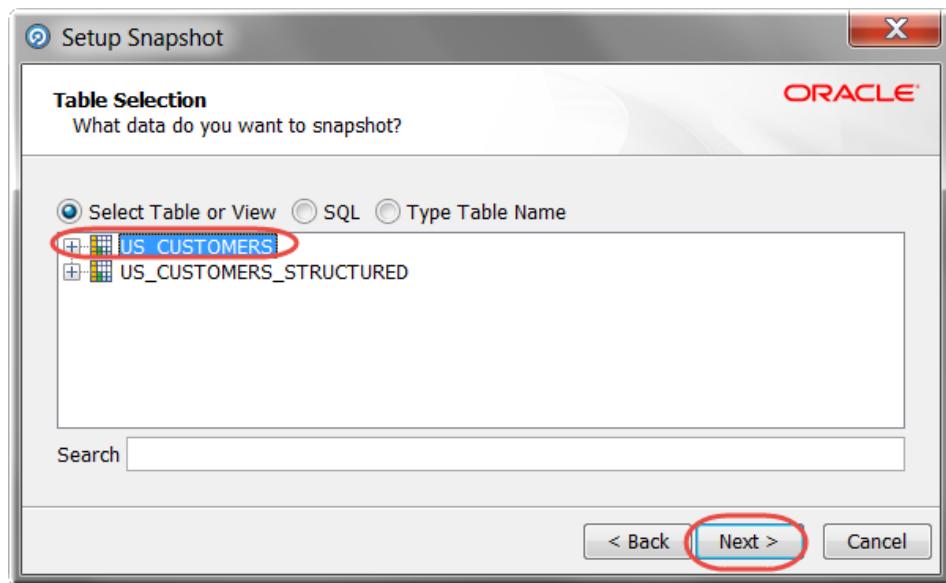


5. Quay lại **Trình duyệt dự án** và nhấp chuột phải **Dữ liệu được phân loại** dưới của bạn **Khám phá dữ liệu khách hàng** dự án và lựa chọn **Ảnh chụp nhanh mới...**

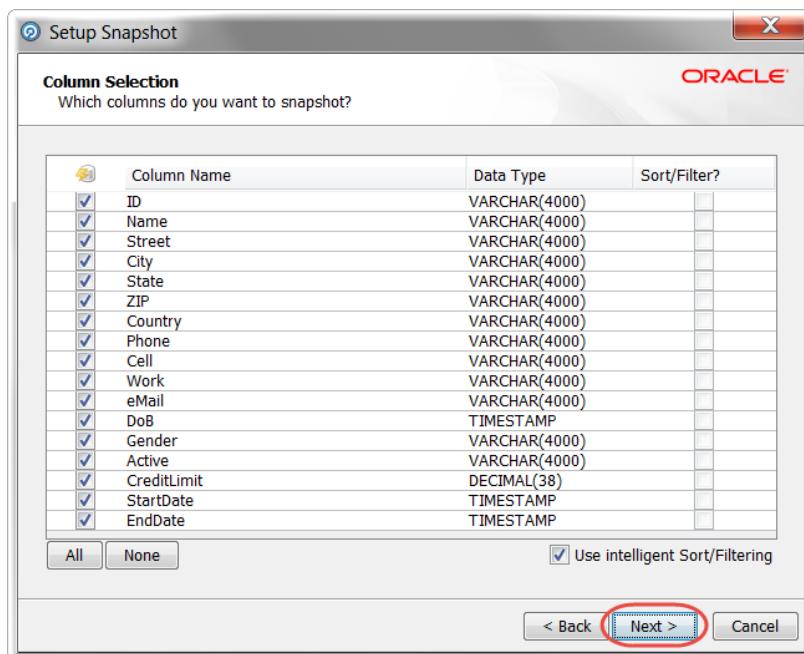


6. Chọn kết nối cơ sở dữ liệu mới được tạo – **Kết nối với cơ sở dữ liệu Oracle**, sau đó nhấp vào **Tiếp theo >** để tiếp tục. Đây là nơi dữ liệu cho ảnh chụp nhanh sẽ đến từ. Đối với **Lựa chọn bảng**, chỉ có một bảng được trình bày trong danh sách và được tự động chọn.

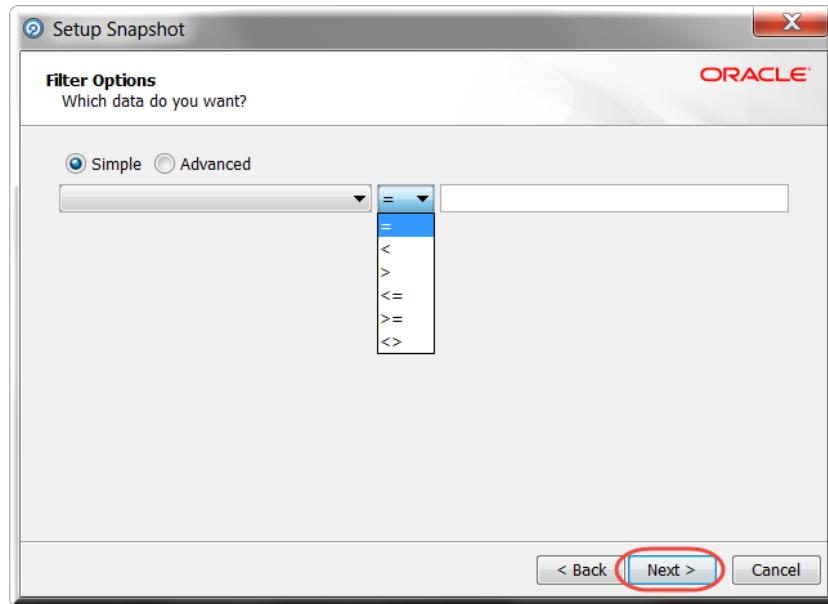
nguồn của Dữ liệu được dàn dựng EDQ cụ thể này là bảng **DỮ LIỆU KHÁCH HÀNG CỦA CHÚNG TÔI**. Giữ lại **Chọn Bảng hoặc Xem** nút radio đã chọn. Nhấp vào **Tiếp theo >** để tiếp tục



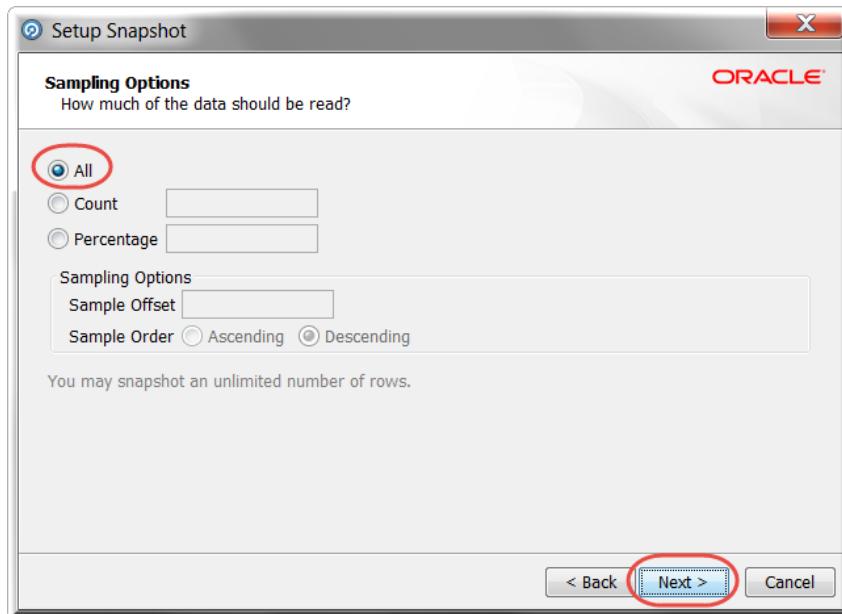
7. Đối với **Lựa chọn cột**, đảm bảo tất cả các cột được chọn để thiết lập ảnh chụp nhanh dữ liệu này, sau đó nhấp vào **Tiếp theo >** để tiếp tục



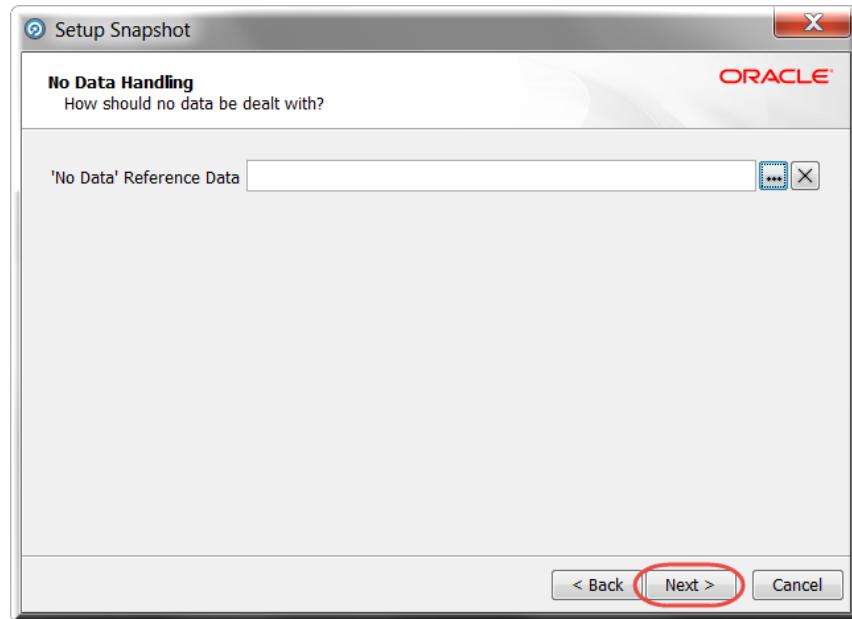
8. Đối với **Tùy chọn bộ lọc**, chú ý **Đơn giản** và **Trình độ cao** các nút radio để lọc dữ liệu theo cột và toán tử. **Trình độ cao** tùy chọn cho phép thêm truy vấn SQL của riêng bạn. Để nguyên giá trị mặc định của **Đơn giản** và nhấp vào **Tiếp theo >** để tiếp tục



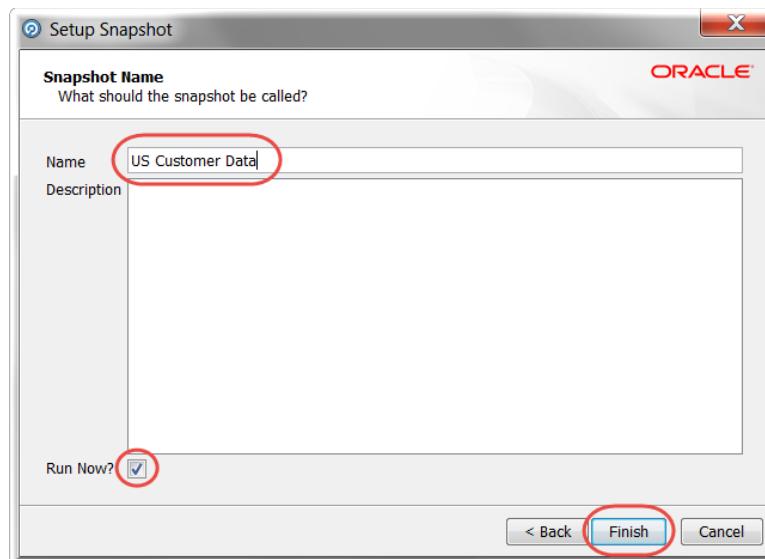
9. **Tùy chọn lấy mẫu** cho phép hành vi của lượng dữ liệu được đọc vào ảnh chụp nhanh thay đổi. Theo mặc định, **Tất cả** nút chọn radio được chọn. Nếu cần, bạn có thể chỉ định một **Đếm** hoặc **Phần trăm** của dữ liệu cần đọc cho ảnh chụp nhanh. Trong ví dụ này, hãy chọn **Tất cả** để biết các tùy chọn lấy mẫu, sau đó nhấp vào **Tiếp theo >** để tiếp tục



10. Để giá trị mặc định trống cho 'Không có dữ liệu' Dữ liệu tham khảo Textfield. Chúng ta sẽ làm việc với Dữ liệu tham chiếu sau trong phòng thí nghiệm khác



11. Đặt tên cho ảnh chụp nhanh của bạn như sau - *Dữ liệu khách hàng Hoa Kỳ*. Đảm bảo **Chạy ngay bây giờ?** Hộp kiểm được chọn, sau đó nhấp vào **Hoàn thành** để hoàn thành và đóng **Ảnh chụp nhanh mới** phù thủy



Lưu ý rằng sau một thời gian trì hoãn ngắn, **Trình duyệt kết quả** được điền dữ liệu có nguồn gốc từ Cơ sở dữ liệu Oracle và lấy từ Ảnh chụp nhanh EDQ. Việc lấy Ảnh chụp nhanh gây ra

Enterprise Data Quality để dàn dựng dữ liệu từ cơ sở dữ liệu vào kho dữ liệu EDQ – nghĩa là một bản sao dữ liệu từ cơ sở dữ liệu được đặt trong kho Enterprise Data Quality. Từ bây giờ, chúng tôi sẽ làm việc với dữ liệu nằm trong US Customer Data Snapshot.

The screenshot shows the Oracle Director application interface. On the left, there is a navigation tree under 'localhost (admin)'. Under 'Customer Data - A Simple Example', several items are listed: 'EDQ 12.1.3 OTN Demonstrations', 'EDQ-CDS', 'EDQ-CDS - Data Quality Health', 'EDQ-CDS - Initialize Reference', 'Exploring Customer Data', 'Data Stores', 'Connection to Oracle Database', 'Staged Data', 'Data & Customer Data', 'Data Interfaces', 'Processors', 'Published Processors', 'Reference Data', 'Results Books', 'External Tasks', 'Notes', 'Watchlist Management', 'Watchlist Screening', 'Reference Data', 'Data Sources', 'Published Processors', and 'Images'. A red box highlights the 'Results Browser' window in the center. This window has a toolbar at the top with icons for search, refresh, and export. Below the toolbar is a grid titled 'Viewing 100 records of 5,438'. The columns are labeled: ID, Name, Street, City, State, ZIP, Country, Phone, Cell, and Work. The data grid contains numerous rows of customer information. At the bottom of the grid, there is a 'Name Filter' input field and a 'Tasks' section.

ID	Name	Street	City	State	ZIP	Country	Phone	Cell	Work
OGA701110	Mrs Christine Hunt	11187 Leadbetter Road	ASHLAND	VA	23005	USA	(703) 636 0289		
SGV498860	Mr Walker Terrell	442 East 144th Avenue	BROOMFIELD	CO	80020	USA	(303) 738 2431	(509) 048 1176	
TBS512904	Mr Lester Young	9010 Meier Road	LAUREL	MD	20723	USA	(301) 465 7444	(344) 719 3600	(453)
MMR493050	Mr. Michael McDaniel	5001 Arden Lane	BAKERSFIELD	CA	93201	USA	(941) 494 2200	(417) 745 6296	
KBA492050	Mrs Phyllis Ford	880 Terminal Place	RICHMOND	VA	23220	USA	(161) 441 6358	(205) 386 7434	(684)
TJC495720	Mrs Marilyn Tan	11847 Shaver Road	SCHOOLCRAFT	MI	49087	USA	(433) 742 9147		
WZ249351	Mrs Mary Hill	502 Oak Street	RED BLUFF	CA	96080	USA	(337) 033 6299	(601) 288 6677	
RVH493200	Mr. Robert Trover	2300 1st Street	GOLDEN	CO	80401	USA	(303) 279 1200		
RAZ491806	Mr. Brian Robles	10800 Foreman	LOWELL	MI	49331	USA	(740) 937 3714		
VVR498569	Mr. Richard Brown	4048 South Lynn Court	INDEPENDENCE	MO	64055	USA	(462) 347 6776	(741) 915 4823	
VERF49469	Mr. Perry Seibert	202 C Street	SAN DIEGO	CA	92101	USA	(644) 938 4573		
VVR498569	Mr. Perry Seibert	4301 Howard Bush Road	NEOSHO	MO	64850-9104	United States	(546) 652 9402	(309) 766 3697	
FNR660911	Mrs Edith Reed	3619 North 33rd Avenue	PHOENIX	AZ	85017	USA	(480) 898 5225		

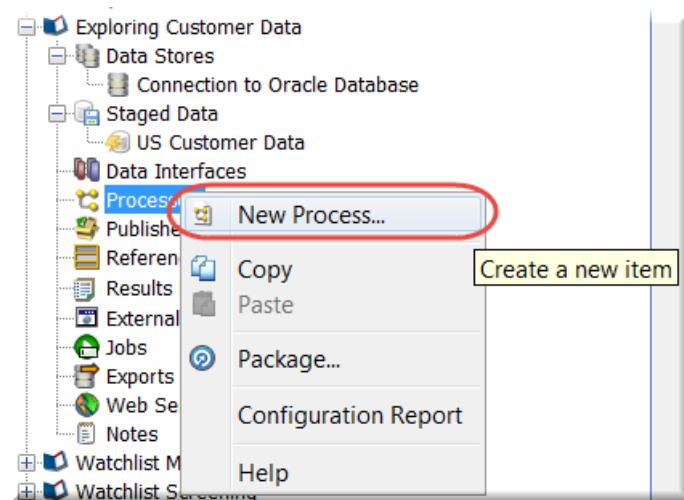
Phòng thí nghiệm 1b: Lập hồ sơ dữ liệu của bạn

Bước đầu tiên để cải thiện chất lượng dữ liệu của bạn là hiểu nó. Bạn biết mình có 'Dữ liệu' – nhưng liệu nó có 'Phù hợp để sử dụng' không? Enterprise Data Quality cho phép người dùng nhanh chóng đánh giá, tìm, điều tra và hiểu các bất thường liên quan đến nội dung dữ liệu, chuẩn hóa, mối quan hệ và trùng lặp trong số những thứ khác. EDQ cho phép người dùng hiểu dữ liệu của họ bằng cách khám phá, làm nổi bật và truyền đạt các bất thường về dữ liệu trong dữ liệu đang được lập hồ sơ.

Như bạn sẽ học, Profiling có thể dẫn đến nhiều hiểu biết khác nhau về nguồn dữ liệu và mục tiêu của bạn bao gồm các giá trị ngoại lệ, giá trị tối thiểu và tối đa, ngày không hợp lệ và tính đầy đủ của bản ghi. Nó cũng có thể hiển thị tần suất xuất hiện của các giá trị cụ thể. Ví dụ, bạn có bao nhiêu giá trị duy nhất trong một trường và một giá trị cụ thể được sao chép bao nhiêu lần? Các chủ đề về profiling này và nhiều chủ đề khác là trọng tâm của phòng thí nghiệm sau đây của chúng tôi: Tạo và chạy quy trình Profiling.

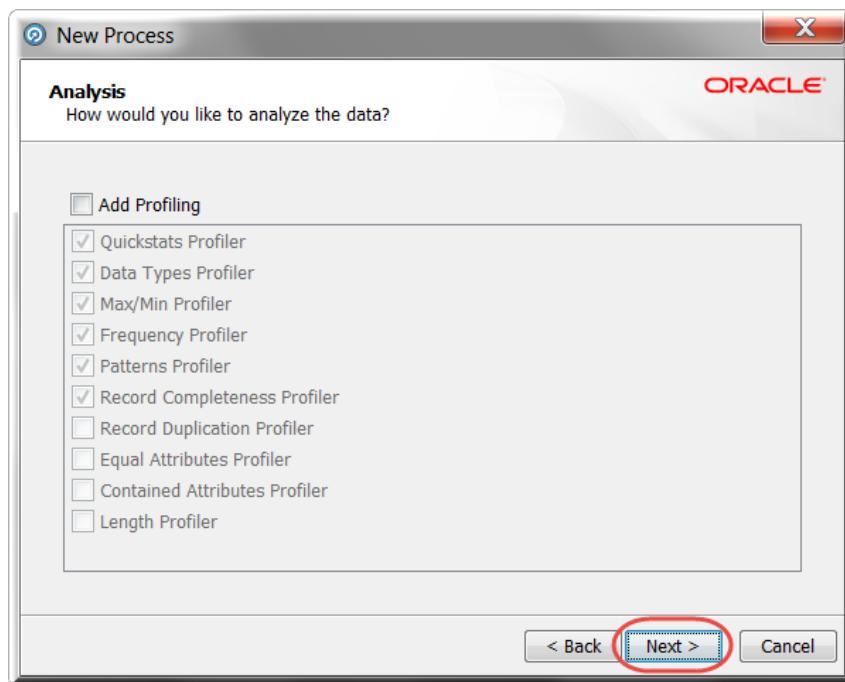
Tạo và chạy quy trình lập hồ sơ

- Điều hướng đến Trình duyệt Dự án và nhấp chuột phải vào **Các quy trình** dưới của bạn **Khám phá dữ liệu khách hàng** Dự án, sau đó nhấp vào **Quy trình mới...**



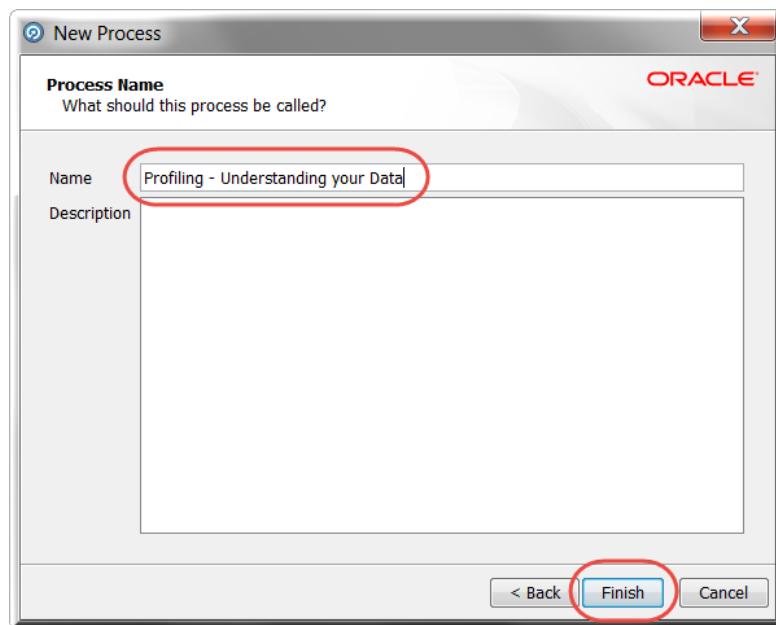
- Chọn Dữ liệu đã được phân đoạn trước đó, **Dữ liệu khách hàng Hoa Kỳ**, sau đó nhấp vào **Tiếp theo >** để tiếp tục

3. Lưu ý rằng bạn có thể tùy chọn chọn **Thêm hồ sơ** trong khi tạo Quy trình mới này. Chúng tôi sẽ thêm bộ xử lý Hồ sơ của riêng mình trong vài bước tiếp theo. Bỏ chọn hộp kiểm chưa kiểm tra và nhấp vào **Tiếp theo >** để tiếp tục



Khi lựa chọn **Thêm hồ sơ** khi tạo một quy trình mới, dữ liệu sẽ được lập hồ sơ cho tất cả các Profiler được chọn. Khi bạn đang tạo một quy trình cho một tập dữ liệu lớn hoặc rộng, bạn nên thêm Profiler riêng sau khi tạo Quy trình vì có thể mất nhiều thời gian để thực thi tất cả các profiler vì chúng là quá trình tạo dữ liệu trung gian và tính toán chuyên sâu

4. Đặt tên cho quá trình này: **Phân tích dữ liệu - Hiểu dữ liệu của bạn** - sau đó nhấp vào **Hoàn thành**

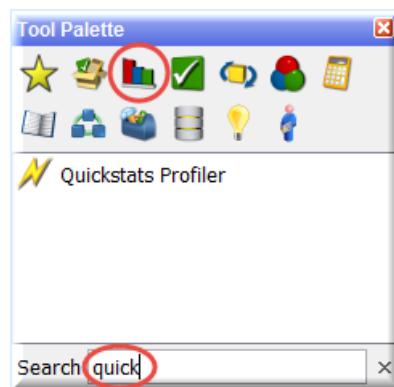


Xin chúc mừng! Một Project Canvas có tab hiện được trình bày với Quy trình mới tạo của bạn. Bạn sẽ lưu ý một **Người đọc** bộ xử lý được tự động thêm vào Project Canvas. Thuật ngữ bộ xử lý sẽ được sử dụng để chỉ các đối tượng được xây dựng sẵn khác nhau được kéo và thả từ **Bảng công cụ**. Tóm lại, mỗi bộ xử lý có thể được cấu hình trong một quy trình để thực hiện một số loại hoạt động trên dữ liệu của bạn. Trong trường hợp này, vì Dữ liệu được phân giai đoạn, **Dữ liệu khách hàng Hoa Kỳ**, đã được chọn trong khi tạo quy trình, một **Người đọc** bộ xử lý tiếp nhận **Dữ liệu khách hàng Hoa Kỳ** đã được thêm vào canvas.

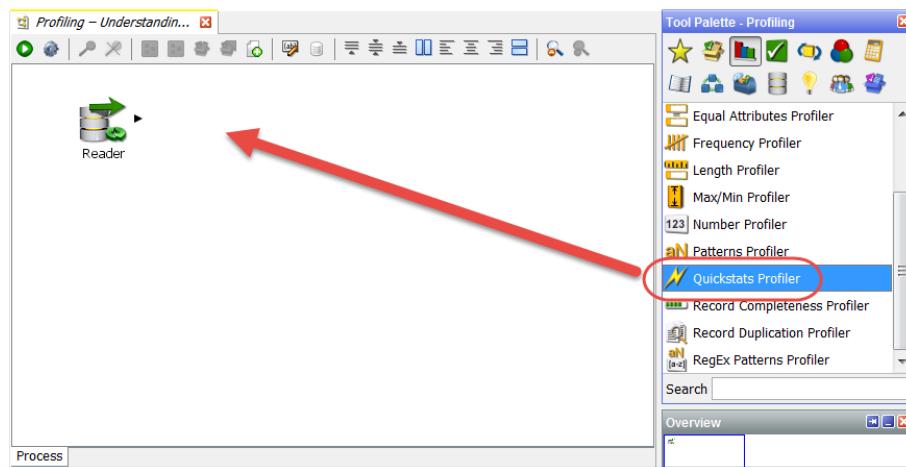
5. Điều hướng đến **Bảng công cụ** và tìm thấy **Hồ sơ** biểu tượng. Tiếp theo, tìm và chọn **Hồ sơ Quickstats** trong họ bộ xử lý Profiling



Sử dụng **Tìm kiếm** hộp bên dưới **Bảng công cụ** cũng cho phép bạn nhanh chóng tìm thấy **Bộ xử lý**.

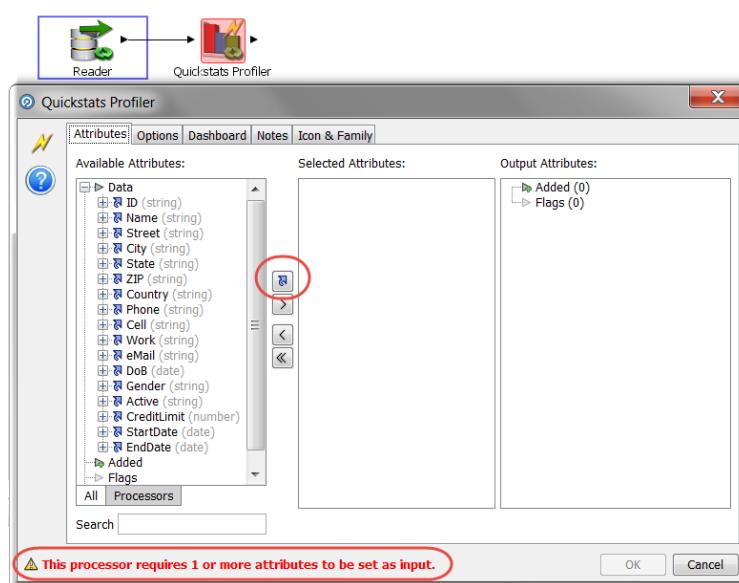


6. Kéo và thả Hồ sơ Quickstats vào Project Canvas



Lưu ý rằng Process Canvas trở nên hoạt động, Process được tạo và Reader và Hồ sơ Quickstats đã được thêm vào. **mũi tên tròn màu xanh lá cây** trên bất kỳ bộ xử lý nào có nghĩa là nó vẫn chưa được thực thi.

7. Di chuột qua hình tam giác đầu ra của Người đọc bộ xử lý. Một mèo công cụ thông tin xuất hiện với tên và mô tả ngắn gọn về bộ xử lý. Nhấp và kéo từ tam giác đầu ra của Người đọc bộ xử lý đến tam giác đầu vào của Hồ sơ Quickstats. Sau khi kết nối và nhả chuột thành công, Hồ sơ Quickstats hộp thoại cấu hình sẽ xuất hiện:



8. Lưu ý thông báo màu đỏ thông báo cho bạn rằng **Bộ xử lý này yêu cầu ít nhất một**

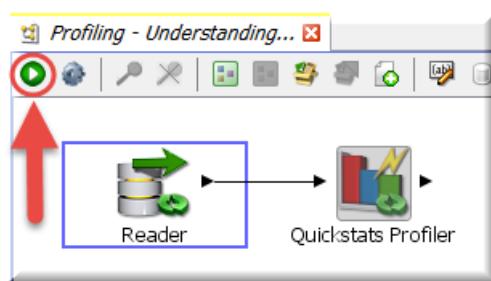
thuộc tính được thiết lập làm đầu vào. Nhấp vào **Chọn tất cả** biểu tượng, như được hiển thị trong ảnh chụp màn hình ở trên. Điều này sẽ chọn tất cả các thuộc tính có sẵn từ **Dữ liệu khách hàng Hoa Kỳ** dữ liệu được dàn dựng. Nhấp vào **ĐƯỢC RỒI** để tiết kiệm



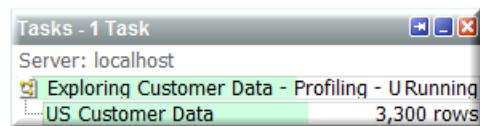
Bộ xử lý mới cũng hiển thị biểu tượng màu xanh lá cây 'chưa chạy'.

Khi một quy trình được chạy, kết quả được lưu trữ trong kho lưu trữ Enterprise Data Quality và các biểu tượng màu xanh lá cây 'chưa chạy' sẽ biến mất. Khi các bộ xử lý tiếp theo được thêm vào, chỉ những bộ xử lý đó cần được chạy miễn là các bộ xử lý khác không thay đổi. Như bạn nhớ lại, khi chúng tôi tạo ra **Quy trình mới**, chúng tôi đã rời khỏi **Thêm Profilers** hộp kiểm không được chọn. Do đó, việc thêm các bộ xử lý riêng biệt sẽ tiết kiệm thời gian và dung lượng đĩa vì chỉ các bộ xử lý Profiler mà chúng ta cần sẽ được thêm vào và chạy thay vì bộ mặc định gồm nhiều bộ xử lý profiler (một số trong số đó chúng ta có thể không cần để tạo Profile dữ liệu được dàn dựng biển **Người đọc**).

9. Quá trình này hiện có một **Người đọc** và một **Hồ sơ Quickstats**. Nhấp vào **Chạy** biểu tượng trên thanh công cụ để chạy quy trình



*Sự tiến triển có thể được quan sát trong **Thanh tác vụ** ở góc dưới bên trái của Director khi tiến trình chạy. Khi tiến trình kết thúc, các biểu tượng 'chưa chạy' sẽ biến mất khỏi canvas để cho thấy bộ xử lý có dữ liệu liên quan đến chúng.*



10. Nhấp vào **Người đọc** bộ xử lý để xem dữ liệu đầu vào thô được lưu trữ trong ảnh chụp nhanh dữ liệu được dàn dựng. Điều này sẽ được hiển thị trong **Trình duyệt kết quả**

11. Tiếp theo nhấp vào **Hồ sơ Quickstats** để xem đầu ra của bộ xử lý

Results Browser

Job: Profiling - Understanding your Data Latest Run: Oct 16, 2015 12:02:18 PM - 12:03:41 PM

Viewing all 17 records

Input Field	Record Total	With Data	Without Data	Singleton	Duplicates	Distinct Values	Comment
ID	5438	5438	0	5438	0	5438	Complete; Possible key
Name	5438	5438	0	5327	111	5380	Complete; Potentially damaged key; Investigate duplicates
Street	5438	5438	0	5319	119	5376	Complete; Potentially damaged key; Investigate duplicates
City	5438	5438	0	396	5042	1232	Complete
State	5438	5438	0	12	5426	65	Complete
ZIP	5438	5436	2	490	4948	1823	Investigate blanks
Country	5438	3641	1797	1	5437	10	
Phone	5438	5422	16	5214	224	5247	Potentially damaged key; Investigate blanks ; Investigate ...
Cell	5438	2350	3088	2346	3092	2349	
Work	5438	1156	4282	1154	4284	1156	
eMail	5438	2531	2907	2325	3113	2429	
DoB	5438	5325	113	3177	2261	3934	Investigate blanks
Gender	5438	4380	1058	0	5438	4	
Active	5438	5124	314	0	5438	5	
CreditLimit	5438	5438	0	0	5438	329	Complete
StartDate	5438	3865	1573	0	5438	38	
EndDate	5438	3865	1573	0	5438	74	

Summary statistics view | Data

Các **Hồ sơ Quickstats** cung cấp số liệu chất lượng cơ bản cho một số hồ sơ hoặc giao dịch, làm nổi bật:

- Các cột khóa của ứng viên
- Tính đầy đủ và dữ liệu bị thiếu
- Trùng lặp
- Tính độc đáo và đa dạng của các giá trị

Đối với mỗi **Trường nhập liệu**, số lượng bản ghi (**Tổng số kỷ lục**), **Với dữ liệu**, **Không có dữ liệu**, **Độc thân**, **Bản sao**, **Và Giá trị riêng biệt** được hiển thị. Những kết quả này có thể được quan sát và điều tra để nhanh chóng tìm ra các bất thường về dữ liệu. Ví dụ, có **4 Giá trị riêng biệt** cho **Giới tính** thuộc tính, khi thực sự chỉ nên có hai: Nam và Nữ. Bạn cũng có thể xem chi tiết bất kỳ văn bản màu xanh nào để xem dữ liệu bên dưới.

12. Nhập vào số **3113** được liệt kê cho **e-mail** dưới **Bản sao** cột trong **Trình duyệt kết quả**.
Nhập vào **Đếm** siêu liên kết cho **e-mail** địa chỉ không chứa nội dung / trống

eMail	Count	%
Richard.L.Smith@xtmail.com	2907	53.5%
Paul.J.Smith@shockmail.com	2	<0.1%
Pamela.N.Arrey@snowmail.com	2	<0.1%
Natalie.L.Simpkins@scene46.com	2	<0.1%
Nancy.J.Fidler@xtmail.com	2	<0.1%
Diane.R.Mccoy@scene46.com	2	<0.1%
Michael.N.Burford@xtmail.com	2	<0.1%
Michael.D.Pate@xtmail.com	2	<0.1%
Michael.D.Aparicio@thu.com	2	<0.1%
David.J.Cooley@xtmail.com	2	<0.1%
Daniel.C.Palermo@scene46.com	2	<0.1%
Lillie.K.Gilbert@scene46.com	2	<0.1%
Kelly.F.Farmer@xtmail.com	2	<0.1%
Judy.D.Mills@scene46.com	2	<0.1%
Karl.A.Jordan@thu.com	2	<0.1%
Julius.S.Gann@xtmail.com	2	<0.1%
Touret.T.Cohen@shockmail.com	2	<0.1%

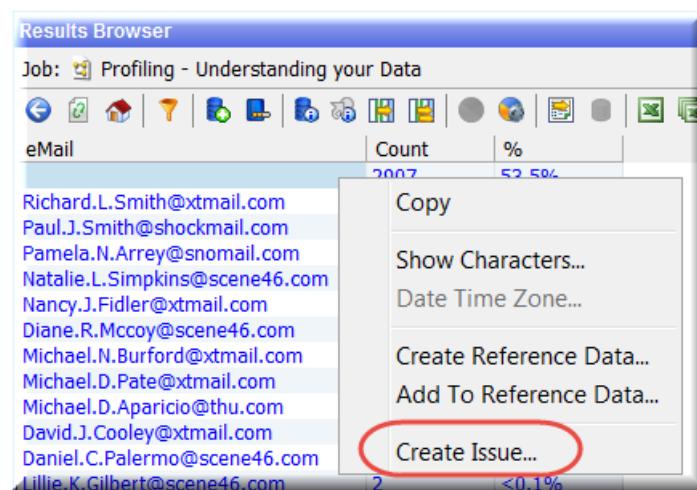
Summary statistics view: "eMa... Data

13. Nhấp vào **Trình duyệt kết quả** để quay lại chế độ xem trước đó của Drill-down trên một trong các giá trị không null. Chúng tôi quan sát thấy có một số giá trị trùng lặp **e-mail** giá trị (**Đếm của 2**) trong hệ thống có thể yêu cầu điều tra thêm từ quan điểm bản ghi trùng lặp

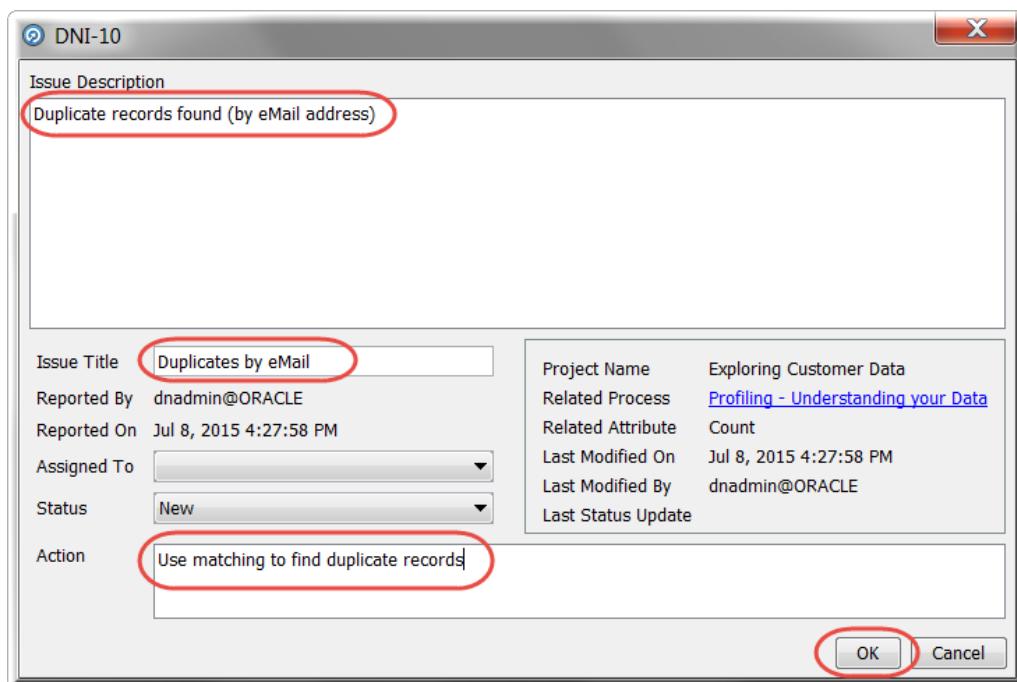
eMail	ID	Name
Susan.J.Winston@shockmail.com	HXP477685	Mrs Susan Winston
Susan.J.Winston@shockmail.com	CSF646812	Mrs Susan Winston

Số lượng lớn các sản phẩm rỗng **e-mail** giá trị (**53,5%**) cũng có thể biểu thị vấn đề 'Dữ liệu phù hợp để sử dụng' tùy thuộc vào yêu cầu/SLA dữ liệu của dữ liệu Khách hàng

14. Enterprise Data Quality có nhiều tính năng để tạo ra một môi trường cộng tác, nơi người dùng có thể nêu vấn đề với DBA, Nhà phát triển ứng dụng, Nhà phát triển ETL, người viết BI Dashboard và Chủ doanh nghiệp khi họ tìm thấy. Chúng tôi sẽ khám phá **Quản lý vấn đề** chi tiết hơn trong phòng thí nghiệm sắp tới: Quản lý sự cố. Để thêm sự cố trùng lặp (**Đếm > 1**) Bản ghi eMail, nhấp chuột phải vào trường siêu liên kết nơi **Đếm** giá trị là **2** trong **Trình duyệt kết quả** và chọn **Tạo vấn đề...**



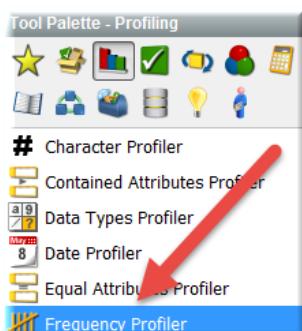
15. Có thể giao vấn đề cho chính bạn (`dnadmin@ORACLE`) hoặc người dùng khác. Bạn cũng có thể nhập nội dung theo dõi **Hoạt động:Sử dụng tính năng khớp để tìm bản ghi trùng lặp**. Vấn đề này cũng bao gồm một liên kết đến chế độ xem quy trình và kết quả nơi vấn đề được tạo ra. Trong **Mô tả vấn đề**, kiểu **Đã tìm thấy bản ghi trùng lặp (theo địa chỉ eMail)**. Nhập vào OK và vấn đề được tạo ra



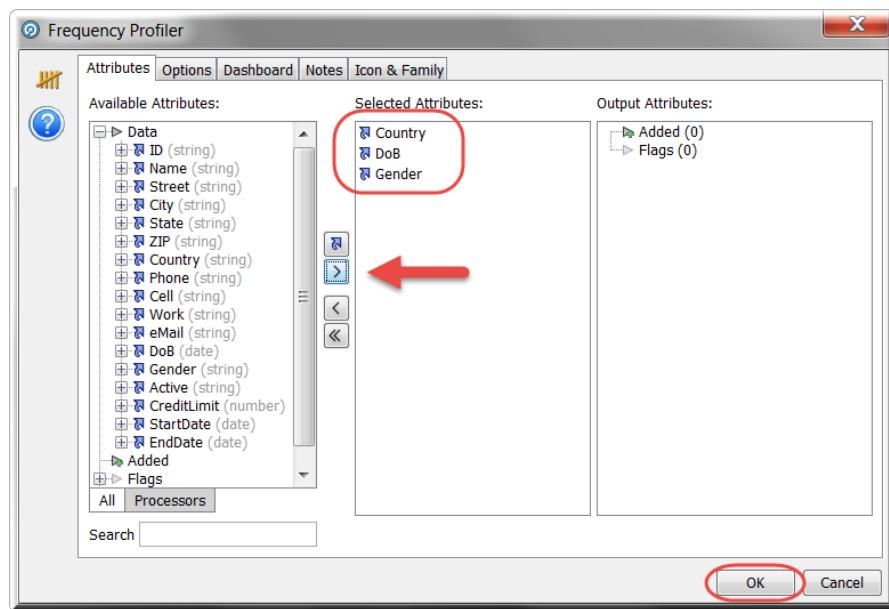
16. Nhấp vào biểu tượng quay lại trong **Trình duyệt kết quả** để quay lại kết quả của **Hồ sơ Quickstats**

Results Browser		
Job:	Profiling - Understanding your Data	
eMa...	Count	%
Richard.L.Smith@xtmail.com	2907	53.5%
Paul.J.Smith@shockmail.com	2	<0.1%
Pamela.N.Arrey@snomail.com	2	<0.1%
Natalie.L.Simpkins@scene46.com	2	<0.1%
Nancy.J.Fidler@xtmail.com	2	<0.1%

17. Quay trở lại **Bảng công cụ - Hồ sơ** và tìm thấy **Tần số Profiler**. Kéo và thả bộ xử lý vào Project Canvas và liên kết tam giác đầu ra của **Hồ sơ Quickstats** vào đầu vào của **Tần số Profiler**



18. Các **Tần số Profiler** hộp thoại xuất hiện. Chọn nhiều **Quốc gia**, **Ngày sinh**, **Và Giới tính** **Thuộc tính có sẵn** và chọn **Thuộc tính**, biểu tượng để thêm các thuộc tính vào **Đã chọn** sau đó nhấn vào **ĐƯỢC RỒI**

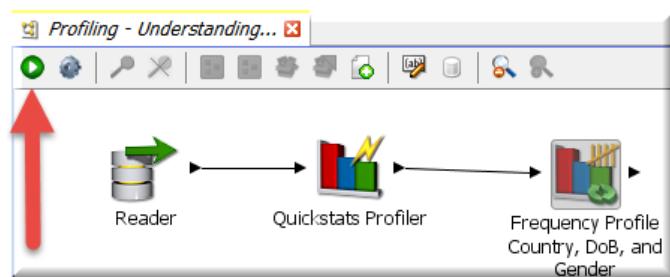


19. Bộ xử lý có thể được đổi tên bằng cách nhấp đúp vào tên bộ xử lý trong khung vẽ. Nhấp đúp vào nhãn hiện có của **Tần số Profiler** và nhập **Hồ sơ tần suất Quốc gia, Ngày sinh và Giới tính** để đổi tên bộ xử lý



Nhấp chuột phải vào biểu tượng bộ xử lý và chọn **Đổi tên** cung cho phép đổi tên bộ xử lý

20. Nhấp vào **Chạy**biểu tượng để bắt đầu quá trình vì Trình phân tích tần suất vẫn chưa chạy và chúng tôi muốn xem kết quả. Đợi quá trình thực hiện hoàn tất



21. Nhấp vào **Hồ sơ tần suất Quốc gia, Ngày sinh và Giới tính**bộ xử lý để xem kết quả trong **Trình duyệt kết quả**. **Ngày sinh, Giới tính**và **Độ tuổi** là những tệp riêng với nhau và không chia sẻ dữ liệu với nhau. Hãy dành một chút thời gian để phân tích những gì các tab khác nhau này nói lên

Results Browser		
Job: Profiling - Understanding your Data		
Value	Count	%
USA	2456	45.2%
	1475	27.1%
US	452	8.3%
U.S.A	344	6.3%
	322	5.9%
United States	228	4.2%
U.S	113	2.1%
Canada	39	0.7%
CAN	8	0.1%
UK	1	<0.1%

Country DoB Gender Data



Có các tab riêng biệt cho từng mục **Thuộc tính được chọn** là một trong nhiều tính năng giao diện người dùng tuyệt vời của EDQ – khả năng sử dụng là một trong 3 điểm khác biệt hàng đầu của EDQ

Bộ xử lý cụ thể này tình cờ cho chúng ta biết rất nhiều về tập dữ liệu của chúng ta chỉ bằng cách quan sát các giá trị khác nhau. Lưu ý có bao nhiêu cách biểu diễn khác nhau của Hoa Kỳ: **Hoa Kỳ, CHÚNG TA, Hoa Kỳ, Hoa Kỳ, VÀ CHÚNG TA**. Sự khác biệt này có thể gây ra các vấn đề lớn với tác giả bảng điều khiển Business Intelligence (BI) và kết quả bảng điều khiển phân tích. Nếu nhóm BI được yêu cầu báo cáo về doanh số bán hàng tại Hoa Kỳ, tác giả bảng điều khiển Business Intelligence phải thực hiện một trong những điều sau (cả hai đều không hấp dẫn):

- Đếm thiếu Doanh số bán hàng của Hoa Kỳ vì anh ấy/cô ấy chọn "Hoa Kỳ" để tìm kiếm và không biết có 5 (và có thể tăng theo thời gian) biểu diễn hợp lệ khác nhau (nhưng tương tự) của Hoa Kỳ trong cột Quốc gia. Cách tiếp cận như vậy dẫn đến việc đếm thiếu Doanh số bán hàng của Hoa Kỳ (chính xác là đếm thiếu 55%).

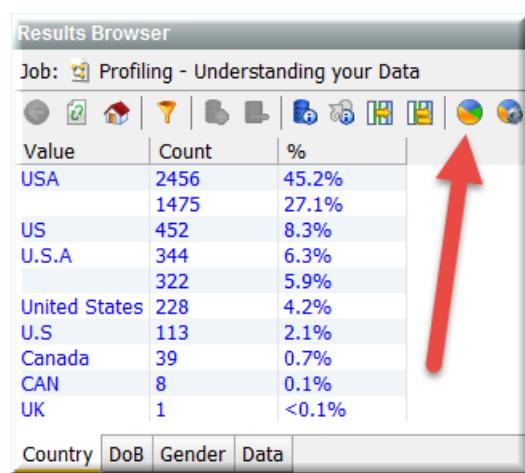
- Phải hỏi "Hoa Kỳ nào?"

- Chịu trách nhiệm xác định tất cả các biến thể của các biểu diễn có thể chấp nhận được của "Hoa Kỳ" trong cột Quốc gia. Và không chỉ cột Quốc gia, tất cả các Cột đều có thể có các vấn đề về nội dung chuẩn hóa – đặt gánh nặng lên các tác giả báo cáo Business Intelligence và bảng điều khiển để tìm ra và xác định tất cả các biến thể có thể chấp nhận được và kết hợp logic đó vào các triển khai báo cáo/bảng điều khiển Business Intelligence. Tại sao lại đưa 'dữ liệu bẩn' vào Data Lake – tại sao không làm sạch nước trước khi đưa vào Lake?

Cuối cùng, có hơn 27% các hàng không có dữ liệu trong cột Quốc gia. Mặc dù chúng ta sẽ không làm như vậy trong Hội thảo này, nhưng người ta có thể dễ dàng tạo Dữ liệu tham chiếu trạng thái và nếu Quốc gia để trống và cột Tiểu bang là một trong 50 tiểu bang hợp lệ – hãy đặt Quốc gia thành giá trị chuẩn hóa là "**Hoa Kỳ**" – thúc đẩy những nỗ lực trong việc chuyển đổi tập dữ liệu Khách hàng từ

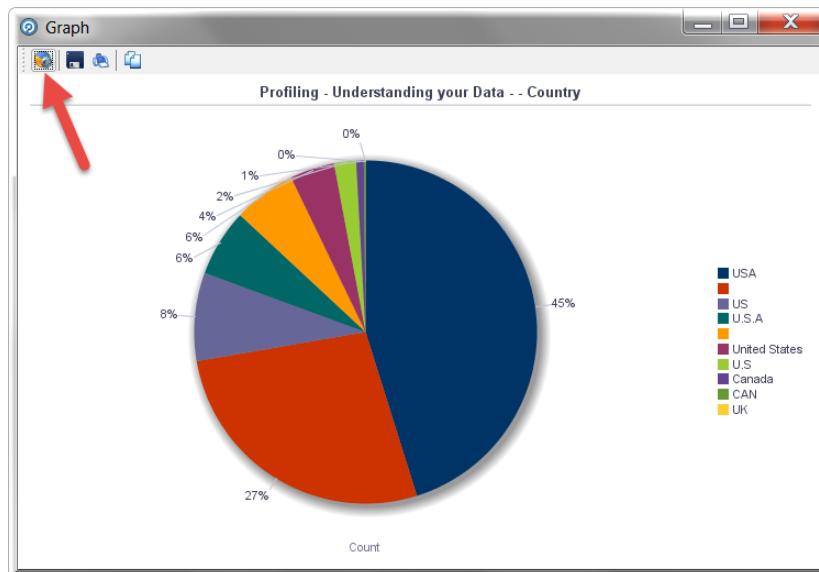
cái gì đó nhiều hơn “chỉ là dữ liệu”, mà là “Dữ liệu phù hợp để sử dụng”. Không chỉ là “Phân tích”, mà là “Phân tích chính xác”.

22. Nhập vào **Kết quả biểu đồ** nút đồ họa  trong **Trình duyệt kết quả** để xem biểu đồ của các giá trị khác nhau được trình bày trong **Đồ thị** hộp thoại



Value	Count	%
USA	2456	45.2%
	1475	27.1%
US	452	8.3%
U.S.A	344	6.3%
	322	5.9%
United States	228	4.2%
U.S	113	2.1%
Canada	39	0.7%
CAN	8	0.1%
UK	1	<0.1%

Country DoB Gender Data

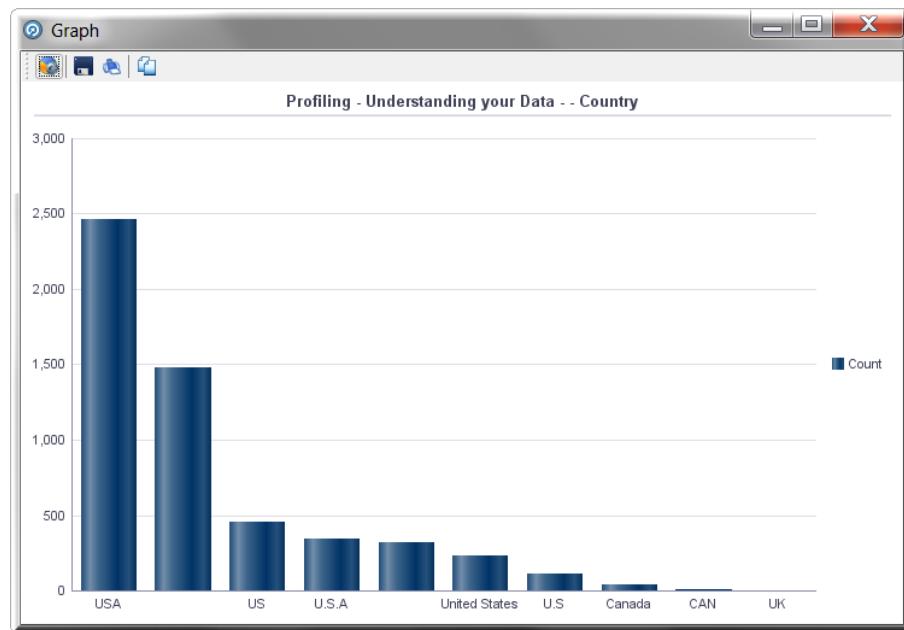
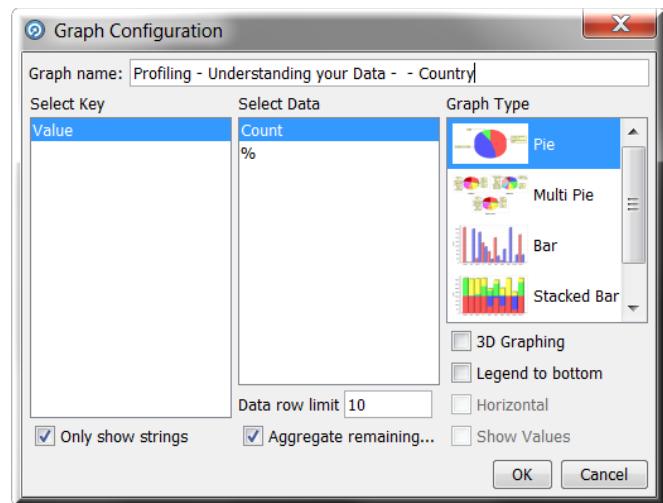


Như chúng ta có thể thấy – mặc dù thực tế là 1% đơn hàng nằm ngoài Hoa Kỳ, hơn một nửa dữ liệu chứa giá trị khác với giá trị phổ biến nhất là **Hoa Kỳ**

23. Nút đầu tiên trên thanh công cụ như được hiển thị trong ảnh chụp màn hình ở trên, sẽ cho phép bạn thay đổi tiêu đề của Biểu đồ, cấu hình Biểu đồ thành một loại hình ảnh khác hoặc sửa đổi loại dữ liệu được hiển thị. Nhấp vào nút này để mở **Cấu hình đồ thị** hộp thoại. Mặc dù khả năng đồ họa trong EDQ không phải là OBI-EE, Cognos hoặc Tableau,

chỉ cần một cú nhấp chuột là có thể hình dung ra **Trình duyệt kết quả**... bạn biết câu tục ngữ cũ – 'một bức tranh có giá trị bằng 1.000 từ' và có giá trị bằng vàng khi cộng tác với những người khác trong nhóm DataMart / Data Warehouse / Data Governance / MDM / Data Scientist và/ hoặc nhóm Giám đốc dữ liệu

24. Chọn **Kiểu biểu đồ thanh**, sau đó nhấp vào **ĐƯỢC RỒI**



25. Ba nút tiếp theo trong biểu đồ cho phép bạn lưu biểu đồ vào tệp hình ảnh, in hoặc sao chép vào bảng tạm để chia sẻ với người khác



26. Đóng lại **Đồ thị** của sổ hộp thoại và nhấp vào **Ngày sinh** tab ở dưới cùng của **Trình duyệt kết quả** để xem kết quả của **Tần số Profiler**

Results Browser

Job: Profiling - Understanding your Data

Value	Count	%
Jan 1, 1970 12:00:00 AM	361	6.6%
	113	2.1%
Jan 1, 1950 12:00:00 AM	58	1.1%
Nov 1, 1975 12:00:00 AM	44	0.8%
Jan 5, 1977 12:00:00 AM	10	0.2%
Jan 1, 1964 12:00:00 AM	6	0.1%
Jan 1, 1954 12:00:00 AM	6	0.1%
Jan 1, 1963 12:00:00 AM	5	<0.1%
Jan 1, 1945 12:00:00 AM	5	<0.1%
Jan 1, 1978 12:00:00 AM	5	<0.1%
Jan 1, 1961 12:00:00 AM	5	<0.1%
Jun 1, 1981 12:00:00 AM	4	<0.1%
Oct 30, 1980 12:00:00 AM	4	<0.1%
Oct 27, 1979 12:00:00 AM	4	<0.1%
Mar 1, 1971 12:00:00 AM	4	<0.1%
Dec 12, 1955 12:00:00 AM	4	<0.1%
Jan 1, 1976 12:00:00 AM	4	<0.1%
May 25, 1981 12:00:00 AM	4	<0.1%

Country Dob Gender Data

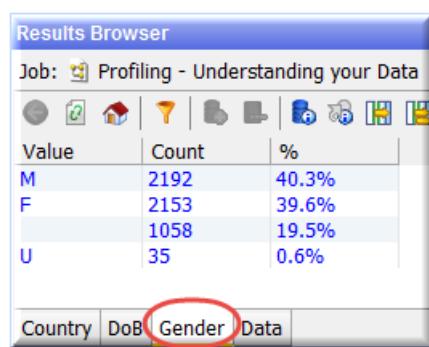
Lưu ý rằng các cột có thể được sắp xếp bằng cách nhấp vào các tiêu đề cột khác nhau (nhấp vào **Đếm** nếu chưa được sắp xếp theo số lượng). Sắp xếp cột trong **Trình duyệt kết quả** có thể là một điều tốt khác

kỹ thuật thăm dò để nhanh chóng xác định các vấn đề bổ sung với dữ liệu của bạn. Ví dụ, bạn sẽ nhận thấy nhiều cá nhân có ngày sinh nhật vào **Ngày 1 tháng 1**. Điều này có thể chỉ ra rằng có một số loại giá trị mặc định được sử dụng với ngày 1 tháng 1 và chất lượng của **Ngày sinh**cột có thể thấp.

Một ví dụ khác, nếu bạn thấy một mục nhập giá trị DoB hiển thị tỷ lệ phần trăm cao bất thường của tổng số hàng (chẳng hạn như **6,6% có Ngày 1 tháng 1 năm 1970**) – điều này sẽ biểu thị một vấn đề ẩn khác với Dữ liệu (có khả năng là giá trị mặc định không mong muốn được hệ thống dữ liệu nguồn sử dụng nếu giá trị là khoảng trống). Tất nhiên, mọi hàng có ngày 1 tháng 1 năm 1970 đều được chèn vào cơ sở dữ liệu nguồn một cách bình thường... nhưng dữ liệu không 'Phù hợp để sử dụng' để phân loại theo độ tuổi hoặc phân tích độ tuổi khác của dữ liệu trừ khi có nỗ lực sửa/làm giàu trường Ngày sinh.

Thiếu 'Phù hợp để sử dụng' của dữ liệu (giá trị mặc định, trống và null cho DoB) giới hạn 'Phù hợp để sử dụng' – giới hạn 'Phù hợp để sử dụng' của dữ liệu vì tác giả Báo cáo/Bảng điều khiển vô tình đưa dữ liệu sai vào kết quả của Bảng điều khiển của họ. Ngoài ra, người dùng Bảng điều khiển mất lòng tin vào Báo cáo/Bảng điều khiển của họ là **nguyên nhân số 1 gây ra thất bại của dự án Data Warehouse / Data Mart / Data Governance (Gartner)**

27. Nhấp vào **Giới tính** tab ở dưới cùng của **Trình duyệt kết quả** để xem kết quả của **Tần số Profiler**

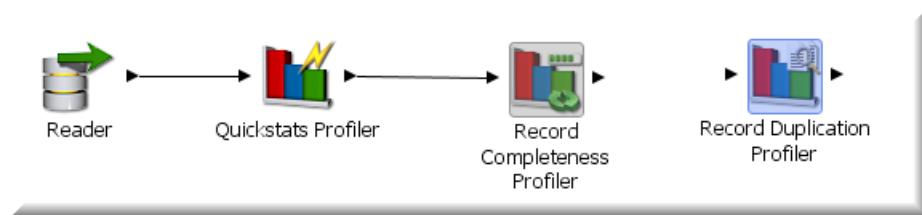


Bây giờ, bạn chắc chắn sẽ thấy rằng tập dữ liệu này cần một số điều chỉnh để có thể sử dụng cho phân tích chính xác. Ví dụ, 19,5% giá trị giới tính ở trên là trống. Trong phòng thí nghiệm sau, chúng tôi sẽ có thể sử dụng bộ xử lý để làm giàu các giá trị giới tính trống để hơn 98% giá trị giới tính là 'Phù hợp để sử dụng' (cho phép phân tích theo Giới tính khi cần)

28. Chúng ta hãy thêm một vài bộ xử lý nữa vào **Phân tích dữ liệu – Hiểu dữ liệu của bạn** quá trình. Quay lại **Bảng công cụ - Hồ sơ** và tìm thấy **Hồ sơ hoàn thiện hồ sơ** và **Trình tạo hồ sơ trùng lặp bản ghi** bằng cách sử dụng hộp Tìm kiếm. Loại **Ghi** trong **Tìm kiếm** trường văn bản để tìm bộ xử lý và kéo thả các bộ xử lý này vào Project Canvas

29. Nhấp và kéo từ tam giác đầu ra của **Hồ sơ Quickstats** bộ xử lý để đường kết nối đạt đến tam giác đầu vào của **Hồ sơ hoàn thiện hồ sơ**

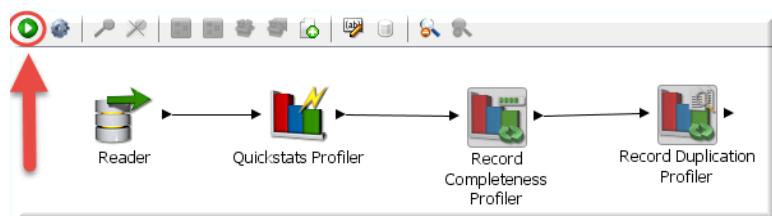
30. Các **Hồ sơ hoàn thiện hồ sơ** hợp thoại cấu hình xuất hiện. Nhấp vào biểu tượng chọn tất cả để tất cả các cột dữ liệu tham gia, sau đó nhấp vào **ĐƯỢC RỒI**



31. Nhấp và kéo hình tam giác đầu ra của **Hồ sơ hoàn thiện hồ sơ** đến tam giác đầu vào của **Trình tạo hồ sơ trùng lặp bản ghi**

32. Các **Trình tạo hồ sơ trùng lặp bản ghi** hợp thoại cấu hình xuất hiện. Nhấp và chọn **Tên thuộc tính** từ **Thuộc tính có sẵn** và nhấp vào biểu tượng để di chuyển thuộc tính đến **Thuộc tính được chọn**. Tương tự, nhấp và chọn **Mã bưu chính** thuộc tính, sau đó nhấp vào **ĐƯỢC RỒI**

33. Nhấp vào **Chạy** (play button icon) biểu tượng trên thanh công cụ để chạy quy trình



Cũng giống như **Hồ sơ Quickstats** cung cấp nhiều chi tiết về tập dữ liệu, **Hồ sơ hoàn thiện hồ sơ** sẽ phân tích các bản ghi với tất cả các thuộc tính đã chọn để hiển thị tính đầy đủ. **Trình tạo hồ sơ trùng lặp bản ghi** sẽ phân tích các bản ghi trùng lặp trên các thuộc tính Tên và Mã bưu chính đã chọn.

34. Nhấp vào **Hồ sơ hoàn thiện hồ sơ** bộ xử lý để xem kết quả trong **Trình duyệt kết quả**. Bạn chỉ có thể thấy rằng **207** của khách hàng trong **Dữ liệu khách hàng Hoa Kỳ** có tất cả **17** trong số **17** thuộc tính đã điền. Nhấp vào **Hiển thị thông tin bổ sung** biểu tượng (info icon). Lưu ý rằng những **207** hồ sơ hoàn chỉnh chỉ tạo nên **3,8%** của toàn bộ tập dữ liệu

Results Browser
Job: Profiling - Understanding your Data

Record Completeness %	Complete Attributes	Matching Records
52.9	9 of 17	2 <0.1%
58.8	10 of 17	43 0.8%
64.7	11 of 17	263 4.8%
70.6	12 of 17	639 11.8%
76.5	13 of 17	1049 19.3%
82.4	14 of 17	1499 27.6%
88.2	15 of 17	1204 22.1%
94.1	16 of 17	532 9.8%
100.0	17 of 17	207 3.8%

Completeness Profile Data

35. Nhấp vào **Trình tạo hồ sơ trùng lặp bản ghi** để xem kết quả. Khoan sâu vào **14** đại diện cho các bản ghi trùng lặp. Chúng ta đã học cách nêu vấn đề trước đó trong phòng thí nghiệm, bạn cũng có thể xuất kết quả sang tệp Excel để gửi cho một cá nhân trong tổ chức để điều tra thêm. Nhấp vào **Xuất sang Excel** biểu tượng trong **Trình duyệt kết quả** thanh công cụ để lưu tệp. (Bạn không cần phải lưu tệp)

Results Browser
Job: Profiling - Understanding your Data

Name	ZIP	ID	Street
Mr Walter Terrell	80020	SGV408860	442 East 144th Avenue
Mrs Tasha Trower	80401	RYW502575	2300 Ford Street
Mr Richard Brown	64055	VVR468569	4048 South Lynn Court
Dr Mudar C Nairne-Clark	33433	FXA588376	6631 Palmetto Circle S
Mr John Morris	22407	BHH630097	8020 River Stone Drive
Mr Ignacio Harris	49444	DJC685684	4800 Harvey Street
Mrs Margaret Johnson	53081	ASW719257	2104 Union Avenue
Mr Walter Terrell	80020	DWC665546	442 East 44th Avenue
Mrs Tasha Trower	80401	KZD465441	2300 Ford Street
Mr Richard Brown	64055	PJB493567	4048 South Lynn Court
Dr Mudar C Nairne-Clark	33433	GTW699731	6631 Palmetto Circle S
Mr John Morris	22407	JHC415718	8020 River Stone Drive
Mr Ignacio Harris	49444	MJV623356	4800 Harvey Street
Mrs Margaret Johnson	53081	NQY460691	2104 Union Avenue



Trong trường hợp có nhiều tab trong Trình duyệt kết quả, bạn có thể nhấp vào nút bên cạnh Xuất sang Excel để **Xuất tất cả các tab sang Excel**.

Trong khi chúng ta có thể tiếp tục thêm Profilers để điều tra thêm dữ liệu Khách hàng Hoa Kỳ, có lẽ tốt nhất là chuyển sang khám phá họ Bộ xử lý EDQ tiếp theo

– **Kiểm toán** – điều này sẽ giúp chúng tôi kiểm tra và chuẩn hóa dữ liệu

Phòng thí nghiệm 2: Kiểm tra dữ liệu của bạn để xác nhận sự khác biệt. Tạo và sử dụng Dữ liệu tham chiếu.

Trong Phòng thí nghiệm này, chúng ta sẽ tập trung vào bộ xử lý Kiểm toán EDQ. Vậy bộ xử lý Kiểm toán là gì và tại sao chúng ta muốn sử dụng chúng? Bộ xử lý kiểm toán, hoặc **kiểm tra**, kiểm tra dữ liệu đầu vào bằng các quy tắc kinh doanh để đánh giá xem dữ liệu đó có phù hợp với mục đích kinh doanh hay không. Các thuộc tính dữ liệu cần kiểm tra thường được xác định bởi:

- Hoạt động lập hồ sơ dữ liệu
- Quy tắc kinh doanh bao gồm một tập hợp các kiểm tra chấp nhận dữ liệu tạo thành SLA (Thỏa thuận mức dịch vụ) chấp nhận dữ liệu giữa hai bên khác nhau - người gửi và người nhận. Tại sao bạn và các đồng nghiệp của bạn lại thực hiện các nhiệm vụ sửa lỗi dữ liệu nếu một tỷ lệ lớn dữ liệu đầu vào từ người gửi được coi là không 'phù hợp để sử dụng'? Chính sách 'trả lại cho người gửi' để sửa dữ liệu có thể là chính sách phù hợp hơn đối với việc không vượt qua các kiểm tra SLA chấp nhận dữ liệu

Bạn sẽ kiểm tra dữ liệu nào (Kiểm toán) để đảm bảo tuân thủ "phù hợp sử dụng" sau Phòng thí nghiệm số một? Quốc gia? Ngày sinh?

Bộ xử lý kiểm toán phân loại từng bản ghi đầu vào để xem nó có **có hiệu lực** hoặc **không hợp lệ** theo như **kiểm tra**. Bộ xử lý kiểm toán cung cấp các luồng dữ liệu đầu ra riêng biệt và dễ truy cập cho các bản ghi hợp lệ và không hợp lệ, cho phép các luồng công việc riêng biệt để xử lý các bản ghi hợp lệ và không hợp lệ trong Quy trình EDQ. Bộ xử lý kiểm toán ngầm sử dụng các quy tắc kinh doanh mà bạn áp dụng cho một thuộc tính dữ liệu nhất định khi lập hồ sơ. Đối với mỗi loại quy tắc kinh doanh mà bạn có thể áp dụng, sẽ có một bộ xử lý Kiểm toán.

Trong Phòng thí nghiệm 2, chúng tôi mở rộng các hoạt động Hồ sơ dữ liệu đã thực hiện trong Phòng thí nghiệm 1 và thực hiện kiểm tra dữ liệu Kiểm toán bằng bộ xử lý Kiểm toán EDQ (các kiểm tra này có thể được coi là "**Tường lửa chất lượng dữ liệu**" cho Kho dữ liệu / Trung tâm dữ liệu của bạn) tận dụng:

- những bất thường về dữ liệu mà chúng tôi phát hiện trong Phòng thí nghiệm 1
- bộ dữ liệu tham chiếu EDQ mở và có thể mở rộng phong phú ('bí quyết' của EDQ)

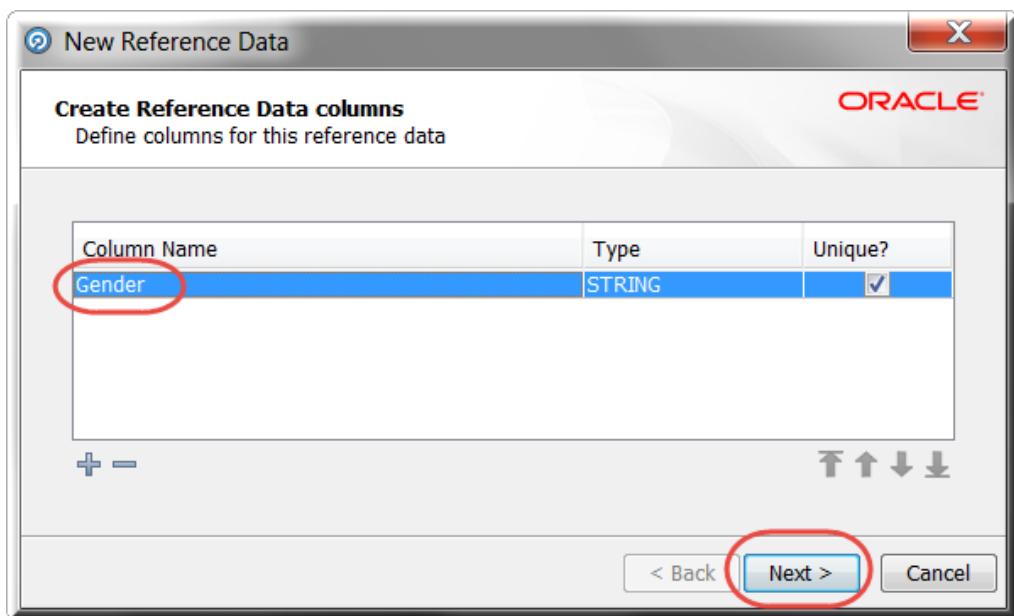
Tạo dữ liệu tham chiếu

1. Đầu tiên, hãy quay lại **Phân tích dữ liệu - Hiểu dữ liệu của bạn** quá trình ở phía bên trái của **Giám đốc**. Bạn có thể đã ở đây rồi, nhưng nếu chưa, hãy tìm dự án của bạn (**Khám phá dữ liệu khách hàng**) trong Trình duyệt Dự án và mở rộng/nhấp đúp vào quy trình Hồ sơ của bạn, **Hiểu dữ liệu của bạn**. Nhấp vào **Hồ sơ tần suất Quốc gia, Ngày sinh và Giới tính**, sau đó nhấp vào **Giới tính** tab ở góc dưới bên trái của **Trình duyệt kết quả**

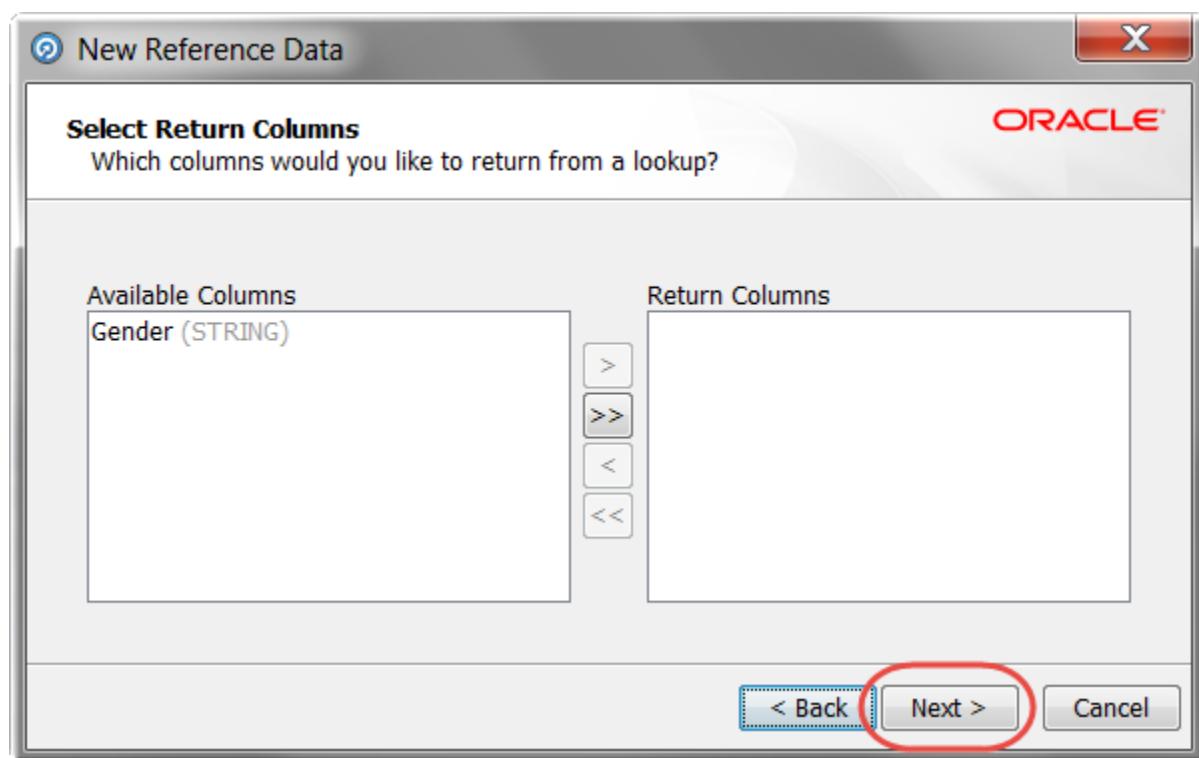
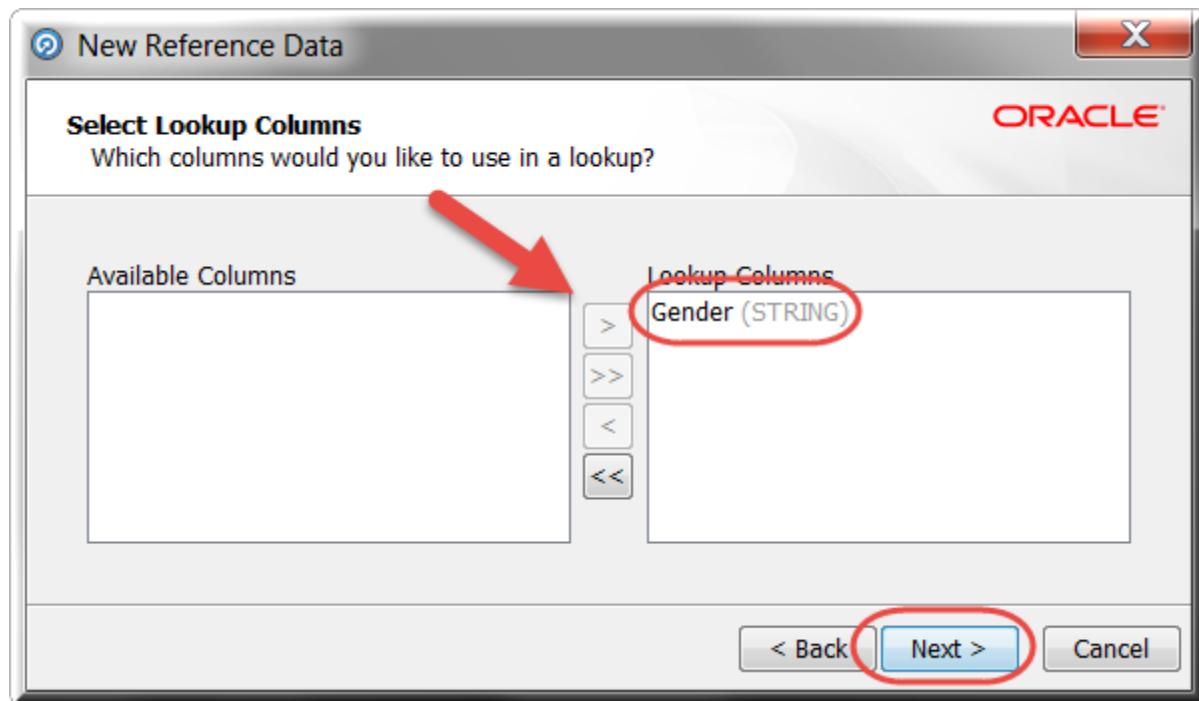
2. Giữ phím CTRL và nhấp vào **Tôi** và **F** giá trị

Value	Count	%
M	2192	40.3%
F	2153	39.6%
	1058	19.5%
U	35	0.6%

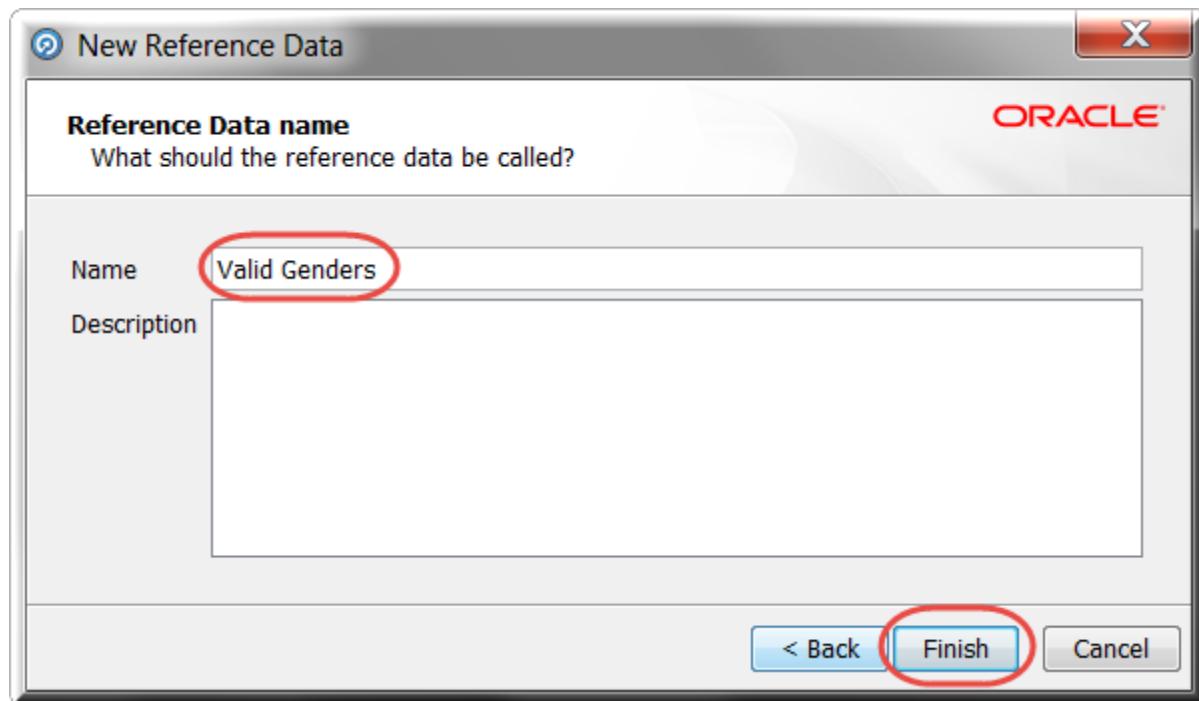
3. Nhấp chuột phải và chọn **Tạo dữ liệu tham chiếu**. Các **Dữ liệu tham chiếu mới** hộp thoại xuất hiện. Đổi tên tên thuộc tính thành **Giới tính**, nhấp chuột **Kế tiếp** để tiếp tục



4. Thêm **Giới tính** đến **Cột tra cứu** sử dụng >nút, sau đó nhấp vào **Tiếp theo >**. Nhấp chuột **Tiếp theo >** trên hai màn hình tiếp theo để giữ nguyên cài đặt mặc định để tiếp tục (chúng tôi sẽ không thêm cột trả về hoặc liên kết (phân loại) dữ liệu tham chiếu này với bất kỳ danh mục nào)



5. Cuối cùng, cung cấp tên cho Dữ liệu tham chiếu này: **Giới tính hợp lệ**, sau đó nhấp vào **Hoàn thành**



6. Các Trình biên tập dữ liệu tham khảo xuất hiện tiếp theo

Tại đây, bạn có thể sửa đổi Dữ liệu tham chiếu để Thêm hàng hoặc Xóa hàng. EDQ cung cấp nhiều loại Dữ liệu tham chiếu khác nhau có sẵn, có thể tăng tốc đáng kể thời gian tạo quy trình kiểm tra dữ liệu. Chúng ta sẽ xem một số ví dụ về Dữ liệu tham chiếu có sẵn trong các phòng thí nghiệm tiếp theo.

7. Tại đây, bạn có thể thấy EDQ đã thu thập các giá trị M và F từ kết quả Hồ sơ của bạn để tạo Dữ liệu tham chiếu tùy chỉnh của riêng bạn. Chúng tôi vừa tạo **nguyên duy nhất** của sự **thật định** nghĩa của một **giới tính hợp lệ**. Nhấp chuột **ĐƯỢC RỒI** để quay lại Project Canvas

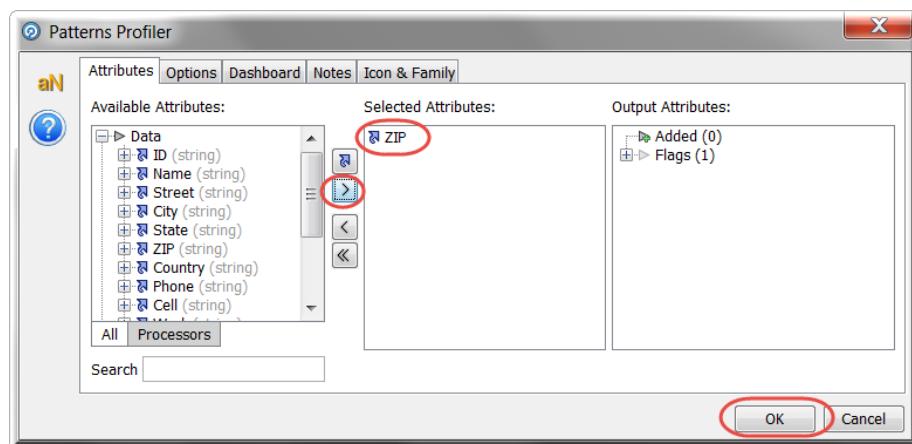
Reference Data Editor - Valid Genders				
Gender	Comment	State	Modified By	Modified On
M	Active	System User	Jul 8, 2015 6:02:...	
F	Active	System User	Jul 8, 2015 6:02:...	

At the bottom of the dialog are several buttons: 'Add Row', 'Add from Clipboard', 'Remove duplicates', 'Delete Rows', 'Delete All Rows', 'OK' (which is circled in red), and 'Cancel'.

Tiếp theo, chúng ta sẽ cần tạo Dữ liệu tham chiếu cho các loại Mã ZIP hợp lệ (nguồn duy nhất của chúng tôi về sự thật của mã ZIP hợp lệ). Trong trường hợp này, chúng ta muốn chuỗi số 5 chữ số hoặc số 5 chữ số theo sau là dấu gạch nối và số 4 chữ số. Để dễ dàng tạo Dữ liệu tham chiếu này, chúng ta cần lập hồ sơ thuộc tính mã ZIP bằng Pattern Profiler.

8. Điều hướng đến **Bảng công cụ** và tìm thấy **Trình tạo hồ sơ mẫu**. Kéo và thả cái này vào **Dự án Canvas** và kéo và thả hình tam giác cuối từ **Trình tạo hồ sơ trùng lặp** bัน ghi đến **Trình tạo hồ sơ mẫu**

9. Các **Trình tạo hồ sơ mẫu** hộp thoại xuất hiện. Chọn **ZIP** từ **Thuộc tính có sẵn** và nhấn > nút để thêm nó vào **Thuộc tính được chọn**, nhấp chuột **ĐƯỢC RỒI** để tiếp tục



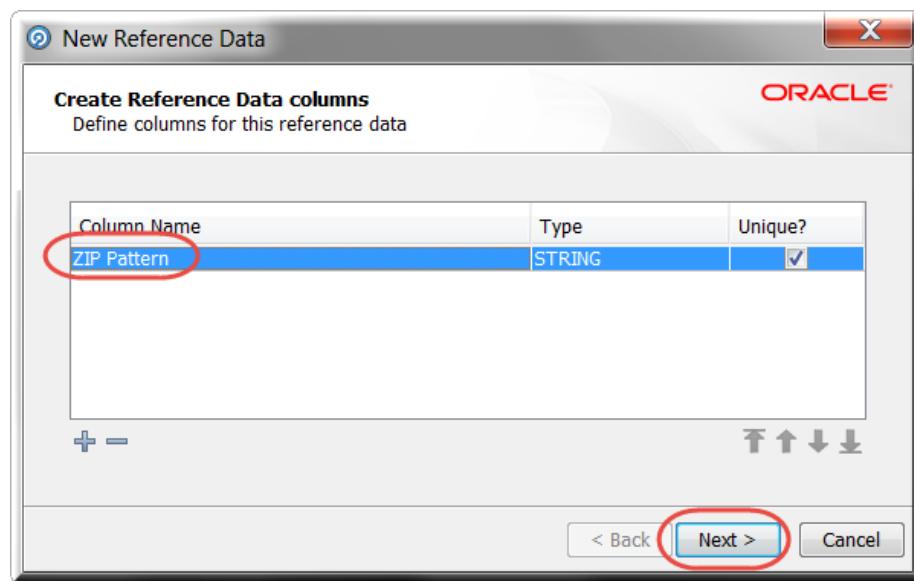
10. Nhấn vào **Chạy** nút trong **Thanh công cụ** ở góc trên bên trái phía trên **Tab Quy trình**. Sau khi quá trình hoàn tất, hãy nhấp vào **Trình tạo hồ sơ mẫu** và xem kết quả trong **Trình duyệt kết quả**

Pattern	Length	Count	%
NNNNN	5	5172	95.1%
NNNNNpNNNN	10	190	3.5%
aNaN_NaN	7	50	0.9%
NNNN	4	16	0.3%
aNaNaNaN	6	4	<0.1%
aaN_Naa	7	3	<0.1%
0	0	2	<0.1%
aNapNaN	7	1	<0.1%

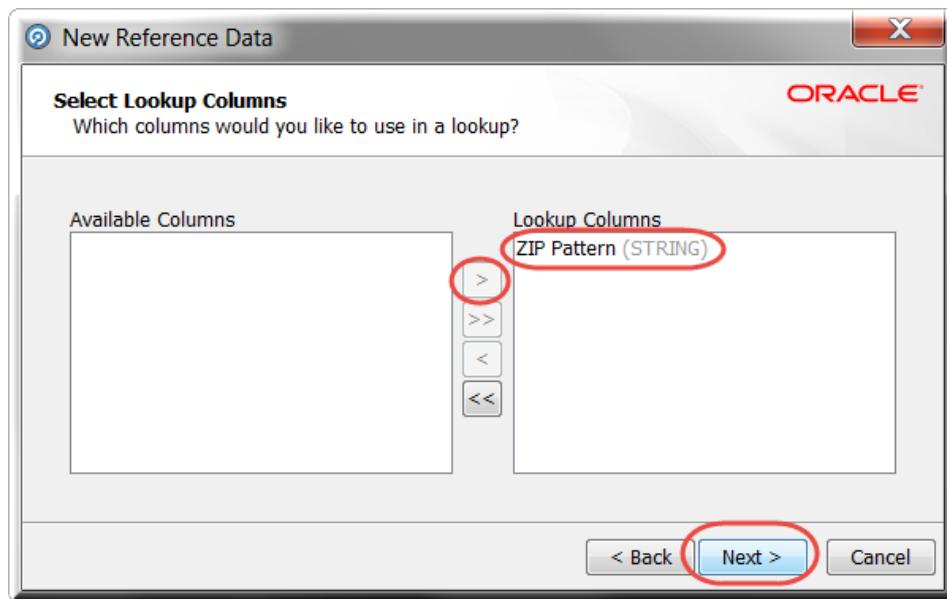
i **N**biểu thị một con số, **P**biểu thị dấu chấm câu, **M**ột biểu thị một ký tự chẵn và **_**biểu thị một khoảng trống.

11. Jetzt haben wir ein Muster für die Referenzdaten definiert. Um das Muster zu speichern, müssen wir es in die Datenbank übertragen. Dazu müssen wir den Wert **NNNNNNV** in das Feld **ZIP Pattern** eingeben und dann auf **Next >** klicken.

12. Nachdem wir das Muster definiert haben, müssen wir es in die Datenbank übertragen. Dazu müssen wir den Wert **NNNNNNV** in das Feld **ZIP Pattern** eingeben und dann auf **Next >** klicken.



13. Fügen Sie das Muster **ZIP** in die Liste der Suchspalten ein, um die Spalte **ZIP** in der Liste der Suchspalten hinzuzufügen. Danach klicken Sie auf **Next >**.

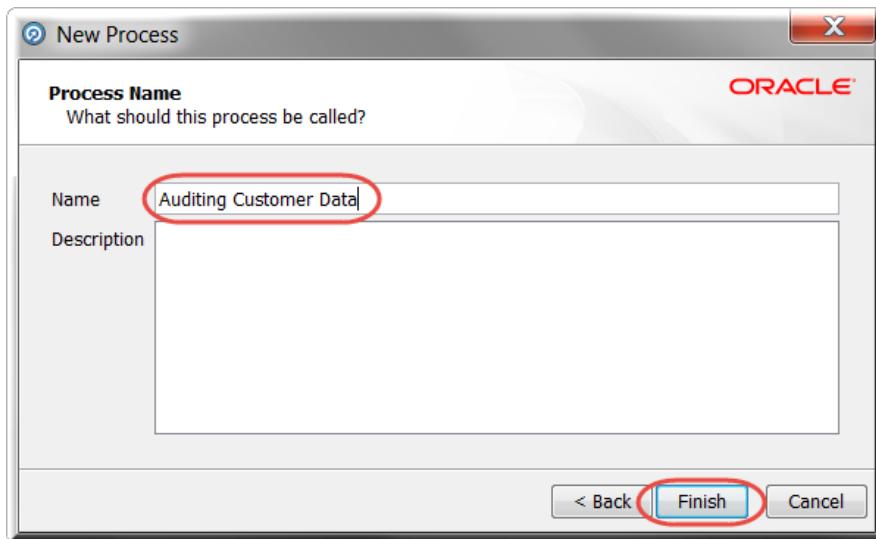


14. Cuối cùng, đặt tên cho Dữ liệu tham chiếu mới này **Mẫu ZIP hợp lệ**, nhấp chuột **Hoàn thành**. Các **Trình biên tập dữ liệu tham khảo** xuất hiện. Nếu bất kỳ giá trị nào được chọn một cách tình cờ, hãy xóa những hàng đó. Nếu không, hãy nhấp vào **ĐƯỢC RỒI** để tiếp tục. Voila! Chúng tôi vừa tạo ra một **nguyên tắc duy nhất của sự thật** định nghĩa cho mẫu mã zip hợp lệ

Bây giờ chúng ta sẽ bắt đầu tạo một Quy trình mới để Kiểm toán Dữ liệu Khách hàng Hoa Kỳ của chúng ta. Dữ liệu Tham chiếu mà chúng ta vừa tạo trong vài bước trước sẽ được một số Bộ xử lý Kiểm toán có sẵn trong Quy trình Kiểm toán (kiểm tra dữ liệu) của chúng ta sử dụng.

Tạo quy trình kiểm toán

15. Quay trở lại **Trình duyệt dự án** ở phía bên trái của cửa sổ Director của bạn và bén dưới **Dự án (Khám phá dữ liệu khách hàng)**, nhấp chuột phải vào **Các quy trình** và nhấp vào **Quy trình mới...**
16. Chọn **Dữ liệu khách hàng Hoa Kỳ**, sau đó nhấp vào **Tiếp theo >**. Nhấp chuột **Kế tiếp** trên màn hình tiếp theo (chúng tôi sẽ không thêm bất kỳ hồ sơ nào nữa ở đây). Đặt tên cho quy trình mới này **Kiểm toán dữ liệu khách hàng**, sau đó nhấp vào **Hoàn thành** để tiếp tục



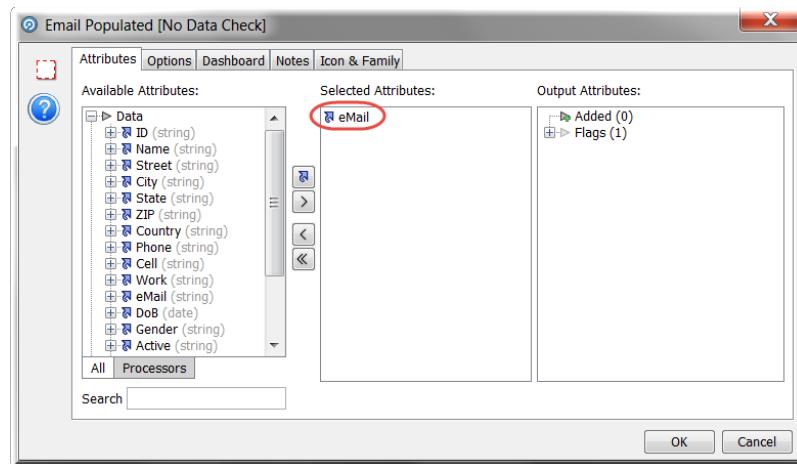
17. Cũng giống như quy trình đầu tiên chúng ta đã tạo, **Bộ xử lý đọc** được tự động thêm vào Dự án Canvas. Điều hướng đến **Bảng công cụ** để tìm Kiểm toán  biểu tượng

Hãy dành chút thời gian để xem xét các Bộ xử lý kiểm toán khác nhau. Di chuột qua các mục nhập khác nhau để có mô tả ngắn gọn.

18. Đầu tiên, kéo và thả một **Không kiểm tra dữ liệu** bộ xử lý vào canvas Process. Nhấp chuột phải vào **Không kiểm tra dữ liệu** bộ xử lý và chọn **Đổi tên** để đổi tên nó thành **Email đã điền** và nhấn phím enter. Kéo và thả hình tam giác cuối từ **Người đọc** với tên mới của bạn **Email đã điền** Quy trình kiểm toán

19. Các **Email đã điền** hộp thoại cấu hình xuất hiện. Chọn **e-mail** từ **Có sẵn**

Thuộc tính và nhấp vào nút để thêm nó vào **Thuộc tính được chọn**, nhấp chuột **ĐƯỢC RỒI** để tiếp tục



20. Nhấp chuột trong **Thanh công cụ** phía trên **Kiểm toán dữ liệu khách hàng** tab quy trình và chọn **Email đã điền** bộ xử lý kiểm toán để xem kết quả

Lưu ý **Không có dữ liệu** và **Với dữ liệu** các giá trị trong **Trình duyệt kết quả**. Nếu muốn, chúng ta có thể tiếp tục phát triển quy trình này bằng cách sử dụng một hoặc nhiều tam giác luồng dữ liệu đầu ra điểm cuối từ Bộ xử lý bằng cách chọn **Dữ liệu**, **Không có dữ liệu** hoặc **Tất cả**.

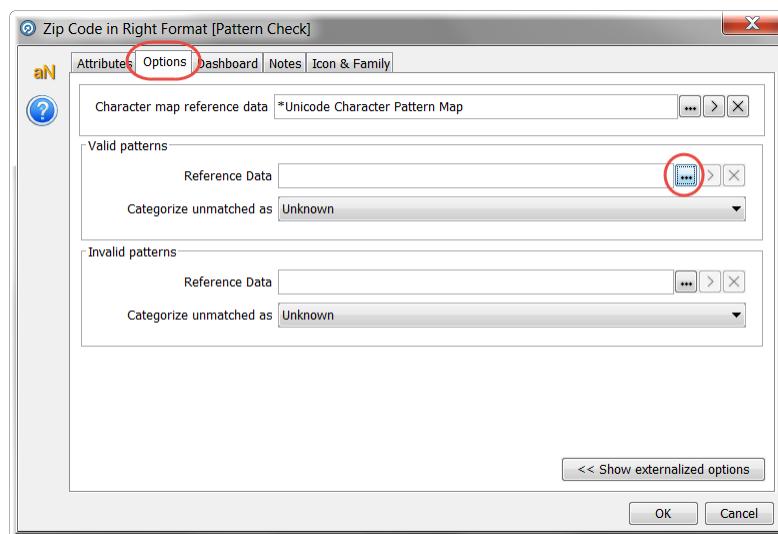
21. Tiếp theo, tìm **Kiểm tra mẫu** Bộ xử lý trong Bảng công cụ. Kéo và thả nó vào khung vẽ và đổi tên thành **Mã bưu chính theo đúng định dạng** bằng cách nhấp chuột phải vào Bộ xử lý kiểm tra mẫu và nhấn phím enter



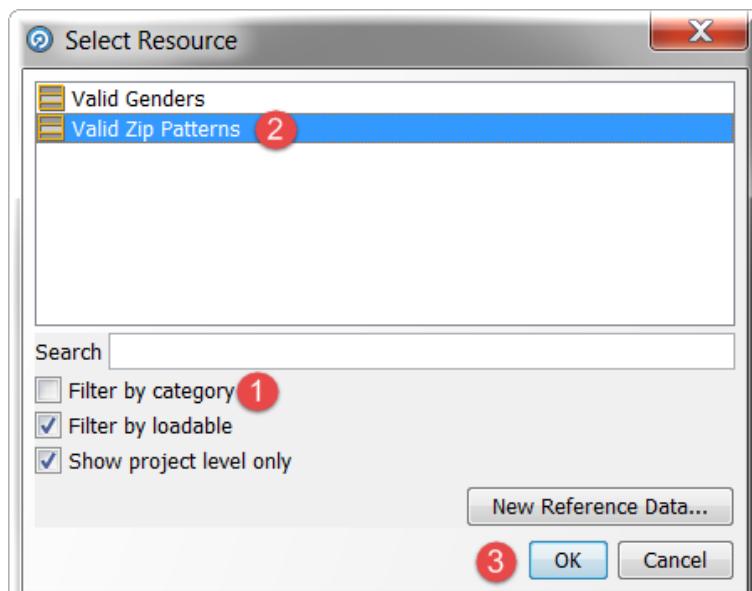
22. Kết nối **Tất cả** kết thúc tam giác từ **Email đã điền đến** **Mã bưu chính theo đúng định dạng** bộ xử lý. Hộp thoại cấu hình cho **Kiểm tra mẫu (Mã bưu chính theo đúng định dạng)** bộ xử lý xuất hiện. Chọn **ZIP** từ **Thuộc tính có sẵn** như là

Trường để xác thực bằng nút 

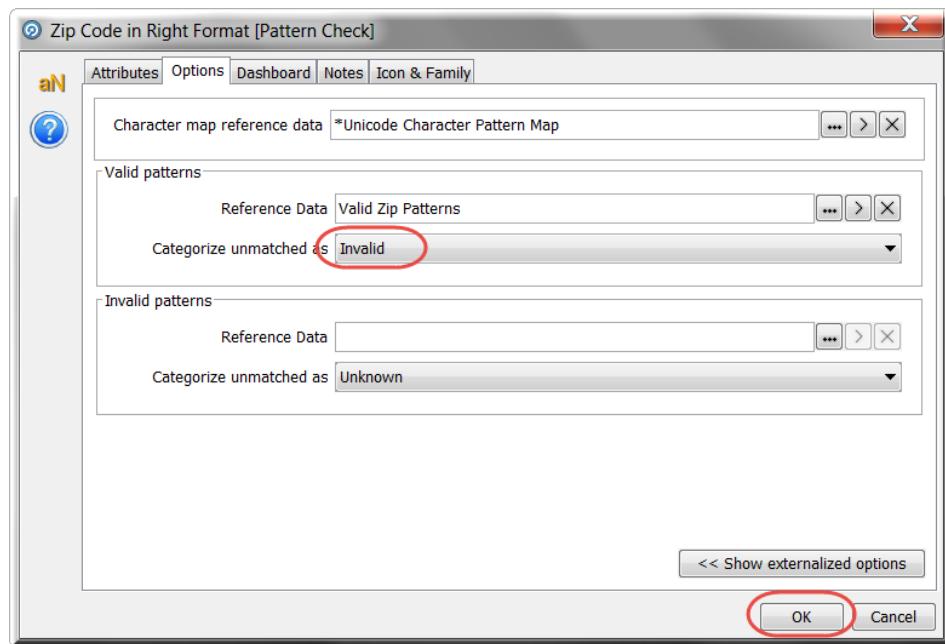
23. Nhập vào **Tùy chọn** tab ở đầu hộp thoại, sau đó nhấp vào **Mẫu hợp lệ**  nút trong **lệp** phần ở giữa cửa sổ



24. Bỏ chọn **Lọc theo danh mục** trong **Chọn tài nguyên** Cửa sổ. Đây là nơi bạn sẽ chọn Dữ liệu tham chiếu mà chúng tôi đã tạo cho các loại Mã bưu chính hợp lệ khác nhau (nguồn duy nhất đáng tin cậy cho định dạng Mã bưu chính hợp lệ) khi bắt đầu phòng thí nghiệm. Nhập vào **Mẫu mã Zip hợp lệ**, sau đó nhấp vào **ĐƯỢC RỒI**



25. Trong phần dưới **Mẫu hợp lệ**, nhấp vào hộp thả xuống để thay đổi **Phân loại không khớp là ĐẾN Không hợp lệ**, sau đó nhấp vào **ĐƯỢC RỒI** để tiếp tục



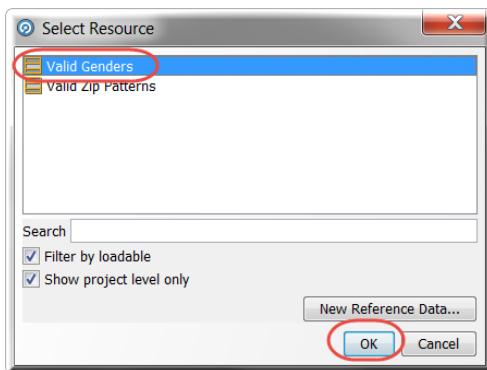
26. Nhấp vào **Chạy** biểu tượng trên thanh công cụ để bắt đầu quá trình. Nhấp vào **Mã bưu chính theo đúng định dạng** để xem kết quả

Lưu ý rằng có **5362 Hồ sơ hợp lệ** và **76 Hồ sơ không hợp lệ**. Nghĩa là có 76 bản ghi không đáp ứng quy tắc 'phù hợp để sử dụng' vì không khớp với NNNNN hoặc NNNNNpNNNN.

27. Quay trở lại **Bảng công cụ** và tìm thấy **Kiểm tra danh sách** bộ xử lý. Kéo và thả nó vào Project Canvas và liên kết **Tất cả** hình tam giác từ **Mã bưu chính theo đúng định dạng** đến **Kiểm tra danh sách** Bộ xử lý

28. Chọn **Giới tính** thuộc tính và thêm vào **Trường để xác thực**. Sau đó nhấp vào **Tùy chọn** tab ở đầu hộp thoại để thêm Dữ liệu tham chiếu cho Giới tính (nguồn duy nhất của định nghĩa sự thật cho giới tính hợp lệ) được tạo ra khi bắt đầu phòng thí nghiệm này. Nhấp vào

biểu tượng ở bên phải của **Dữ liệu tham khảo** trường văn bản trong phần dành cho **Giá trị hợp lệ** để duyệt Dữ liệu tham chiếu và chọn **Giới tính hợp lệ** Tham chiếu dữ liệu, sau đó nhấp vào **ĐƯỢC RỒI** để tiếp tục



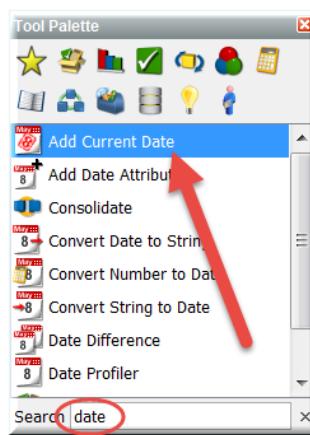
29. Nhấp chuột **ĐƯỢC RỒI** để đóng **Kiểm tra danh sách** hộp thoại. Nhấp đúp vào **Kiểm tra danh sách** bộ xử lý để đổi tên nó thành **Kiểm tra giới tính hợp lệ**. Cuối cùng, nhấp vào **Chạy** nút trên thanh công cụ để bắt đầu quá trình



Một số thông tin này đã được trình bày trong phòng thí nghiệm Profiling; tuy nhiên, các bộ xử lý này cho phép bạn chuyển sang sử dụng các loại bộ xử lý khác để làm việc với *Tất cả*, *Có hiệu lực*, *Không xác định* và/hoặc *Không hợp lệ* kết quả

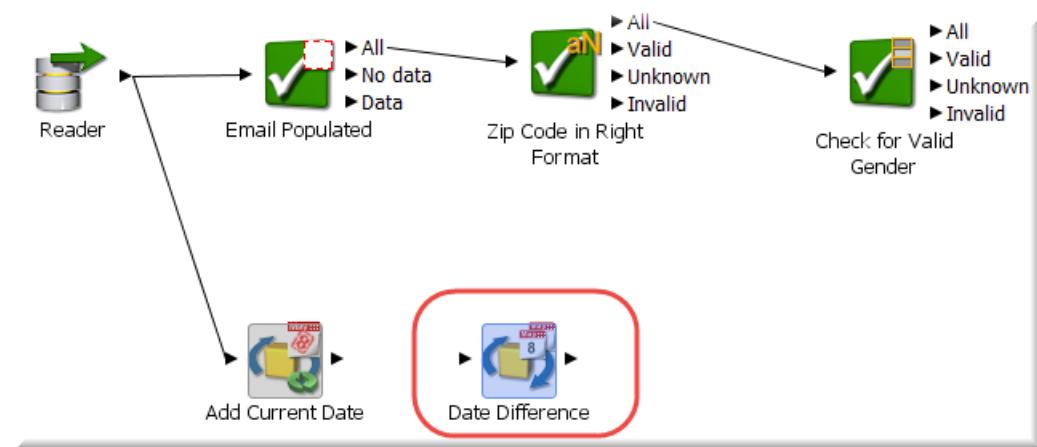
Trong ví dụ kiểm toán tiếp theo này, chúng tôi sẽ kiểm tra Bộ dữ liệu khách hàng Hoa Kỳ của mình đối với những cá nhân dưới 18 tuổi và ghi những bản ghi đó vào Dữ liệu theo giai đoạn. Vì Tuổi không được bao gồm trong bộ dữ liệu, chúng tôi sẽ tính toán bằng cách sử dụng cột Ngày sinh và ngày hiện tại.

30. Để bắt đầu, hãy quay lại **Bảng công cụ** và tìm kiếm **Thêm ngày hiện tại** bộ xử lý. Kéo và thả nó vào Project Canvas (bạn có thể tìm thấy bộ xử lý bằng cách nhập **ngày** vào trường văn bản tìm kiếm của Bảng công cụ)

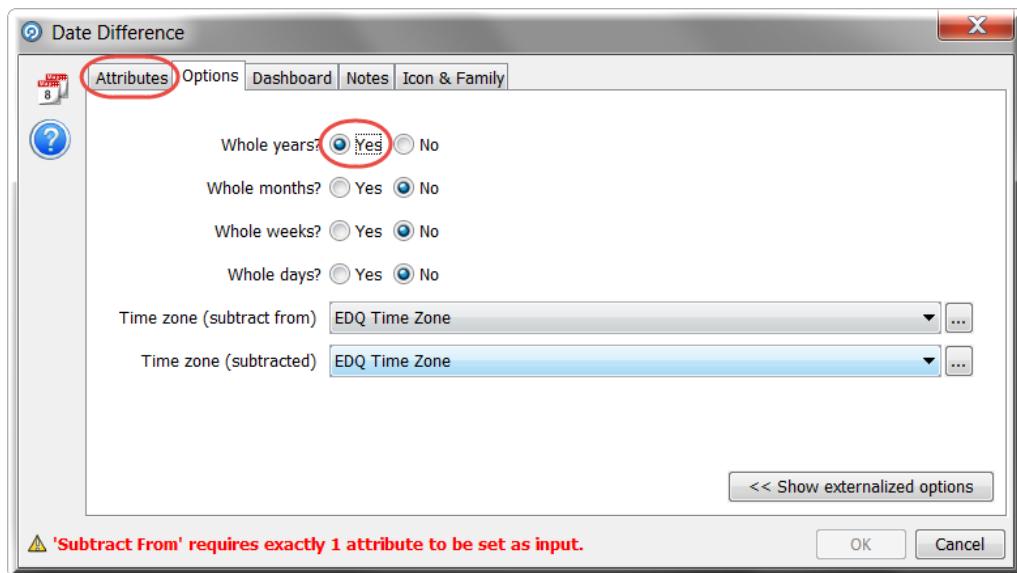


31. Kết nối tam giác cuối từ **Người đọc** đến tam giác đầu vào của **Thêm ngày hiện tại** Bộ xử lý. Không cần cấu hình nào khác cho bộ xử lý này, hãy nhập vào **ĐƯỢC RỒI**
32. Quay trở lại **Bảng công cụ** để tìm thấy **Ngày Chênh Lệch** Bộ xử lý và kéo thả nó vào Project Canvas. Kết nối hình tam giác cuối từ **Thêm ngày hiện tại** đến tam giác đầu vào của **Ngày Chênh Lệch** bộ xử lý

Các **Ngày Chênh Lệch** và **Thêm dữ liệu hiện tại** bộ xử lý sẽ được sử dụng để tính Tuổi của mỗi Khách hàng. Đây cũng là một ví dụ về cách EDQ có thể (và thường thực hiện) thêm các cột siêu dữ liệu bổ sung để sử dụng bởi nhiều bộ xử lý hạ lưu khác nhau trong một Quy trình EDQ



33. Nhập vào **Tùy chọn** tab ở đầu hộp thoại Chênh lệch ngày. Chọn **Đúng** tùy chọn nút radio cho **Cả năm sao?**



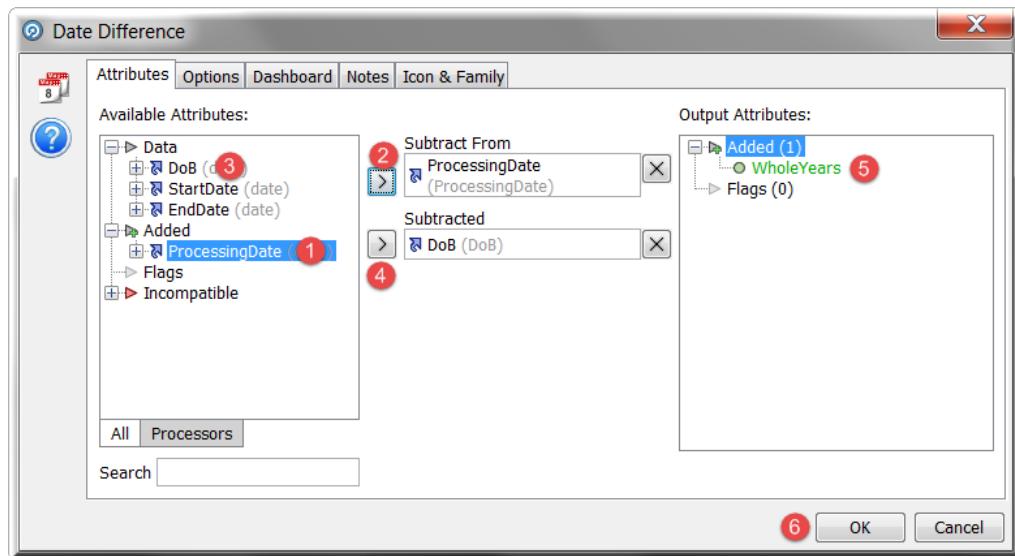
i Nếu bạn muốn tính toán sự khác biệt về Tuổi (hoặc Ngày) theo Tháng, Tuần hoặc Ngày, bạn có thể thực hiện bằng cách chọn các tùy chọn đó. Ngoài ra, vì cả hai **Ngày sinh** và ngày được thêm vào với **Thêm ngày hiện tại** bộ xử lý tình cờ năm trong cùng Múi giờ, chúng tôi sẽ giữ nguyên hai mục thả xuống cuối cùng.

34. Quay trở lại **Thuộc tính** tab bằng cách nhấp vào nó ở góc trên bên trái của Ngày chênh lệch hộp thoại. Thêm **Ngày xử lý** đến **Trừ Từ** tùy chọn sử dụng nút 

35. Thêm **Ngày sinh** đến **Đã trừ** tùy chọn sử dụng tab  nút, sau đó nhấp vào **Tùy chọn** ở đầu hộp thoại Chênh lệch ngày

i Thuộc tính đầu ra hiện đã được thêm vào bên phải hộp thoại. Một ví dụ về EDQ cung cấp siêu dữ liệu bổ sung có thể hỗ trợ chuyển đổi dữ liệu thành 'phù hợp để sử dụng'

36. Nhấp đúp **Toàn Năm** trong vòng **Thuộc tính đầu ra** ở phía bên phải của **Ngày Chênh Lệch** hộp thoại và đổi tên nó thành **Tuổi**, nhấp chuột **ĐƯỢC RỒI** để tiếp tục



37. Nhấp vào **Chạy** nút ở góc trên bên trái của Project Canvas để chạy quy trình.
Sau đó, sắp xếp các bản ghi trong Trình duyệt kết quả bằng cách nhấp vào **Tuổi**cột

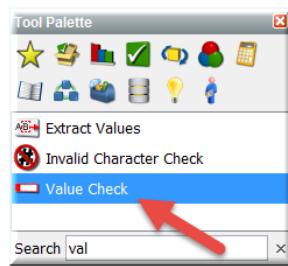
Lưu ý rằng thực sự có những Khách hàng dưới 18 tuổi. Hiện tại chỉ có 2 có thẻ nhìn thấy, nhưng nếu bạn chú ý đến phía bên phải của Trình duyệt kết quả, bạn sẽ thấy chỉ có 100 trong số 1.099 bản ghi mẫu được hiển thị trong tổng số 5.438 bản ghi.

Results Browser		
Job: Auditing Customer Data		
ProcessingDate	DoB	Age
Oct 16, 2015 12:40:41 PM	Nov 28, 1999 12:00:00 AM	15
Oct 16, 2015 12:40:41 PM	Jul 1, 1999 12:00:00 AM	16
Oct 16, 2015 12:40:41 PM	Apr 1, 1996 12:00:00 AM	19
Oct 16, 2015 12:40:41 PM	Jul 13, 1995 12:00:00 AM	20
Oct 16, 2015 12:40:41 PM	Jul 10, 1994 12:00:00 AM	21
Oct 16, 2015 12:40:41 PM	Jul 16, 1990 12:00:00 AM	25
Oct 16, 2015 12:40:41 PM	Nov 25, 1985 12:00:00 AM	29
Oct 16, 2015 12:40:41 PM	Oct 25, 1985 12:00:00 AM	29

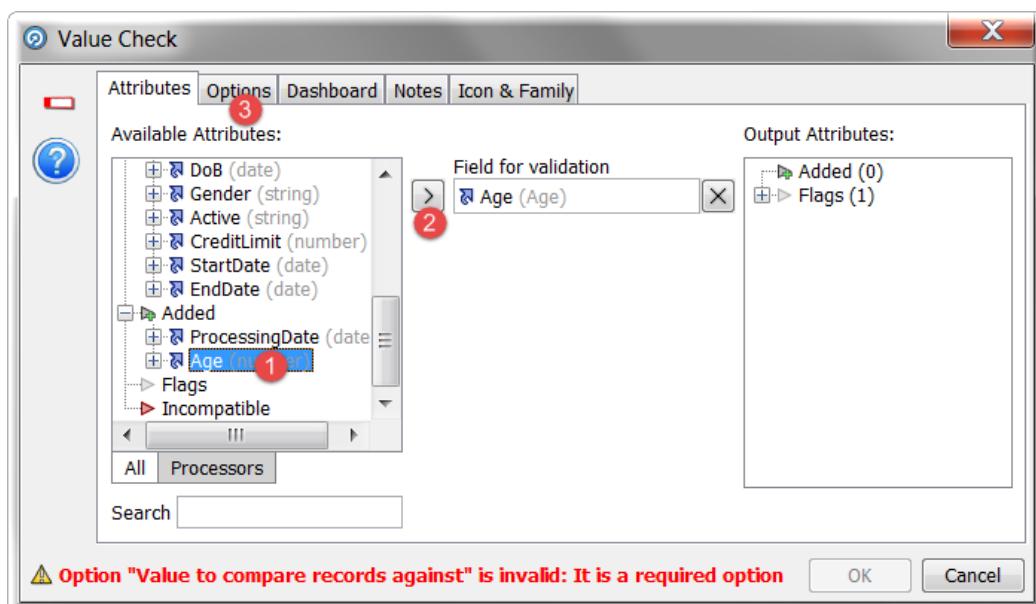
Results Browser										
Job: Auditing Customer Data										
ProcessingDate	DoB	Age	ID	Name	Street	City	State	ZIP	Country	Latest Run: Oct 16, 2015 12:40:39 PM - 12:40:45 PM
Oct 16, 2015 12:40:41 PM	Nov 28, 1999 12:00:00 AM	15	YVF675849	Mr Jason Quintero	2287 Premier Row	ORLANDO	FL	32809	USA	
Oct 16, 2015 12:40:41 PM	Jul 1, 1999 12:00:00 AM	16	ZVQ725272	Mrs Kimberly Collins	200 N 5th Street	GARLAND	TX	75040	US	
Oct 16, 2015 12:40:41 PM	Apr 1, 1996 12:00:00 AM	19	SVO461058	Mrs Jill Gaskill	7070 S 108th Street	LA VISTA	NE	68128	USA	
Oct 16, 2015 12:40:41 PM	Jul 13, 1995 12:00:00 AM	20	GDH521164	Mrs Jessica Turk	46 North 1st Street	FULTON	NY	13069	USA	
Oct 16, 2015 12:40:41 PM	Jul 10, 1994 12:00:00 AM	21	NJZ413152	Mr Phillip Baker	4201 Coldwater Road	FORT WAYNE	IN	46805	USA	
Oct 16, 2015 12:40:41 PM	Jul 16, 1990 12:00:00 AM	25	PUS580434	Mr David Messenger	5837 Distribution Drive	MEMPHIS	TN	38141	USA	
Oct 16, 2015 12:40:41 PM	Nov 25, 1985 12:00:00 AM	29	WAC450327	Mr Charles Holbert	15799 North 1000th Street	EFFINGHAM	IL	62401	USA	
Oct 16, 2015 12:40:41 PM	Oct 25, 1985 12:00:00 AM	29	PLC686322	Mrs Zella Carr	10142 Old Oregon Trail	REDDING	CA	96003		

38. Quay trở lại **Bảng công cụ để tìm thấy **Kiểm tra giá trị** bộ xử lý. Kéo và thả **Kiểm tra giá trị** bộ xử lý cho Project Canvas**

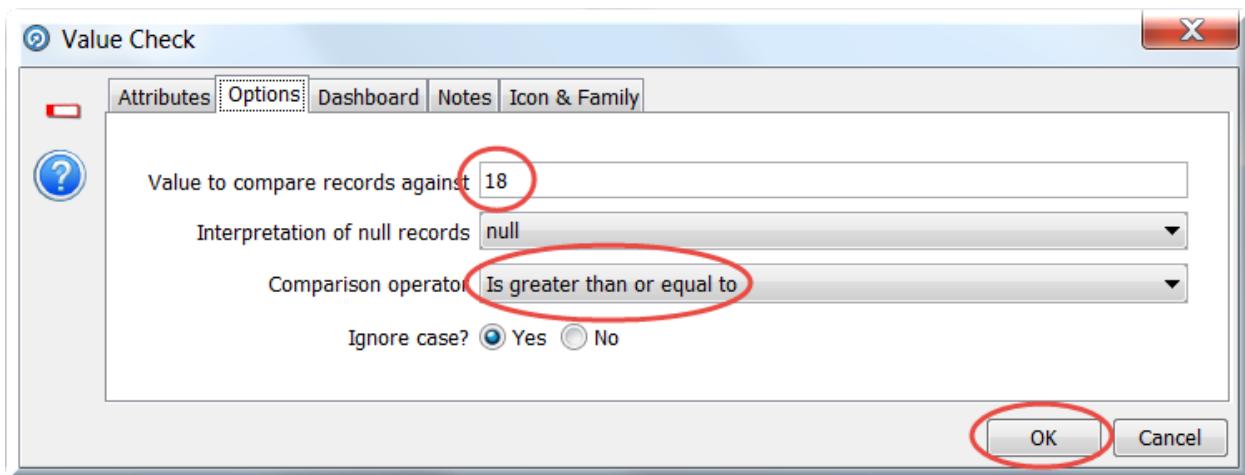
Quá trình này sẽ được sử dụng để tìm các giá trị lớn hơn hoặc bằng 18. Sau đó, chúng tôi sẽ liên kết các bản ghi bị lỗi từ Bộ xử lý này đến một Trình ghi.



39. Kéo hình tam giác cuối từ **Ngày Chênh Lệch bộ xử lý đến tam giác đầu vào của **Kiểm tra giá trị** bộ xử lý. Chọn **Tuổi** thuộc tính như **Trường để xác thực**, sau đó nhấp vào **Tùy chọn** tab ở trên cùng của **Hộp thoại kiểm tra giá trị** để tiếp tục**

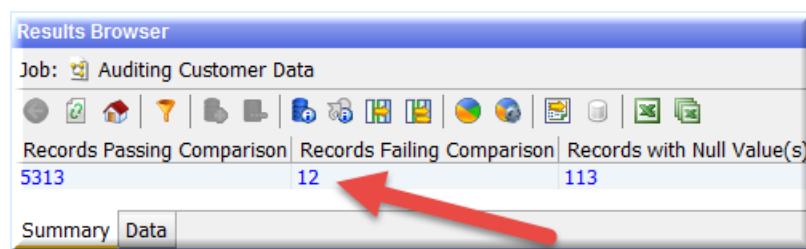


40. Nhập **18 như là **Giá trị để so sánh hồ sơ**. Thay đổi toán tử so sánh thành **Lớn hơn hoặc bằng**, sau đó nhấp vào **ĐƯỢC RỒI** để tiếp tục**



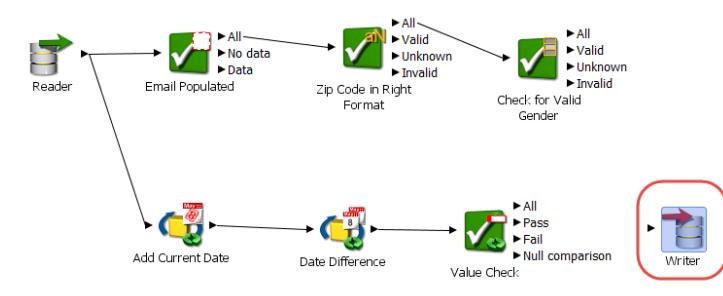
41. Nhấp chuột ở góc trên bên trái của Project Canvas để chạy quy trình

Lưu ý rằng có 12 khách hàng dưới 18 tuổi, trong khi 5313 đáp ứng tiêu chí từ 18 tuổi trở lên.



Trước đó trong phòng thí nghiệm, chúng ta đã thấy cách xuất bản ghi từ Trình duyệt kết quả sang tệp Excel. Chúng ta cũng có thể kéo một hoặc nhiều **Nhà văn** Bộ xử lý cho Project Canvas để ghi kết quả trở lại kho dữ liệu.

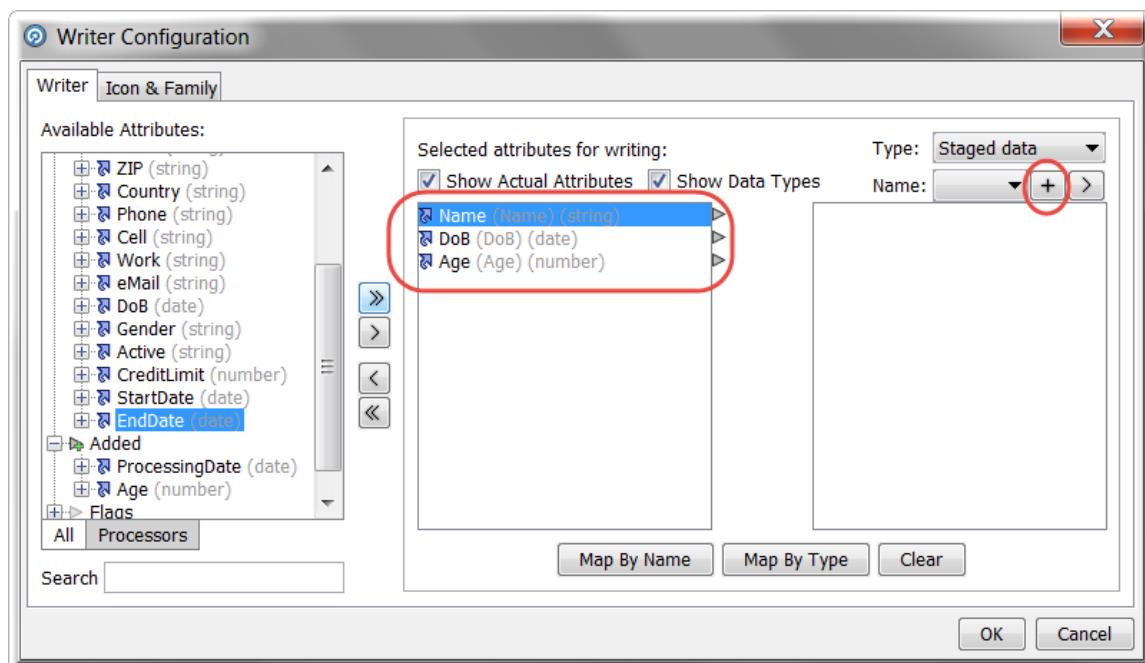
42. Quay trở lại **Bảng công cụ** để tìm kiếm **Nhà văn** và kéo và thả nó vào Project Canvas



43. Kết nối **Thất bại** tam giác cuối của **Kiểm tra giá trị** Bộ xử lý đến tam giác đầu vào của **Nhà văn**

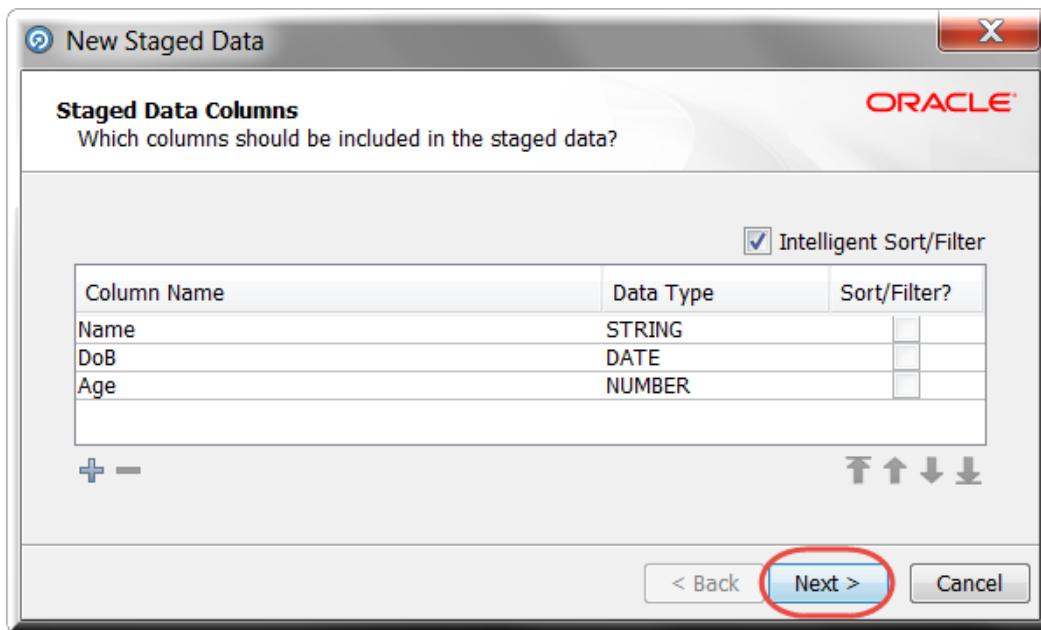
Điều này sẽ cho phép chúng tôi xuất dữ liệu về những Khách hàng dưới 18 tuổi để sử dụng trong các tình huống khác.

44. Chọn nhiều **Tên, Ngày sinh, Vào Tuổi** Thuộc tính từ **Thuộc tính có sẵn** và di chuyển chúng đến **Thuộc tính được chọn để viết** khu vực danh sách. Sau đó nhấp vào + bên dưới **Loại: Dữ liệu được phân loại** thả xuống để tạo một cái mới **Dữ liệu được dàn dựng** đối tượng vì một đối tượng chưa tồn tại

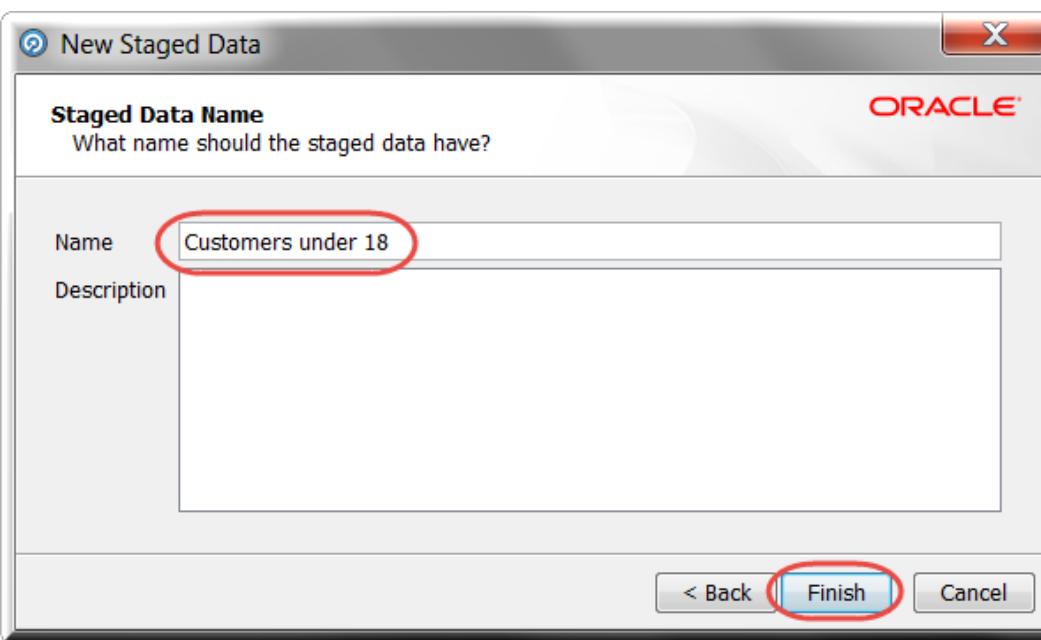


45. Hộp thoại đầu tiên khi tạo **Dữ liệu phân đoạn mới** sẽ nhắc bạn chọn các cột để đưa vào.

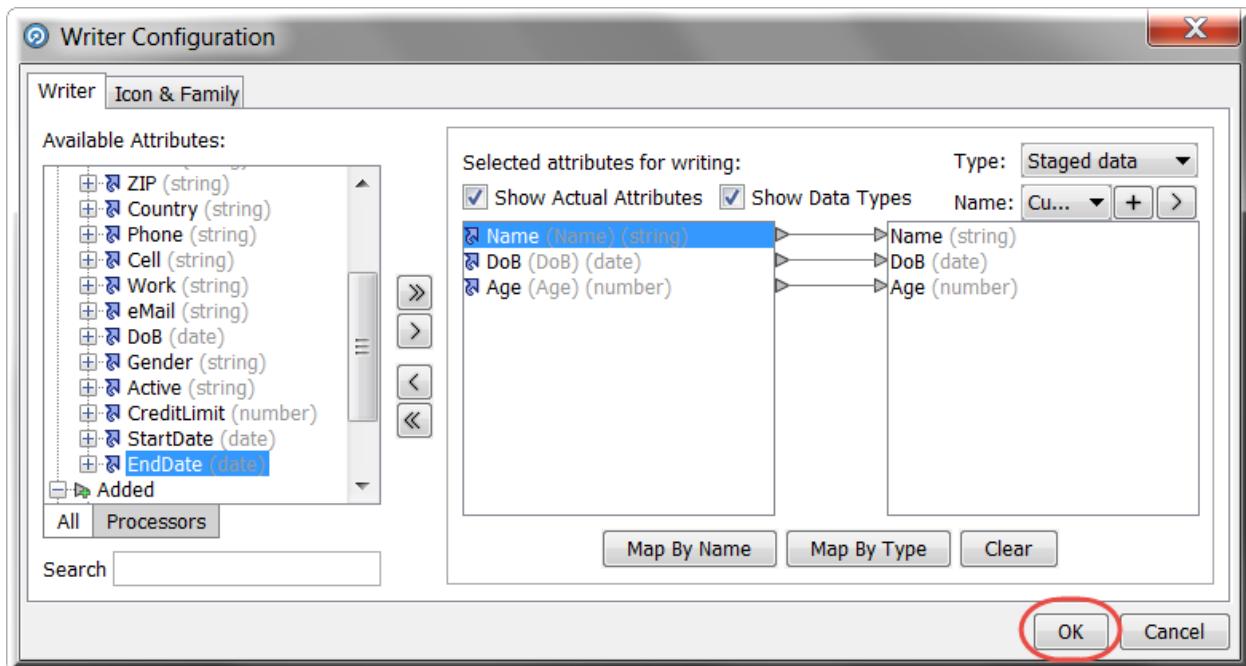
Trong trường hợp này, chúng tôi sẽ giữ nguyên mặc định, nhấp vào **Tiếp theo >** để tiếp tục



46. Đặt tên cho tập dữ liệu được phân giai đoạn này: **Khách hàng dưới 18 tuổi**, sau đó nhấp vào **Hoàn thành**

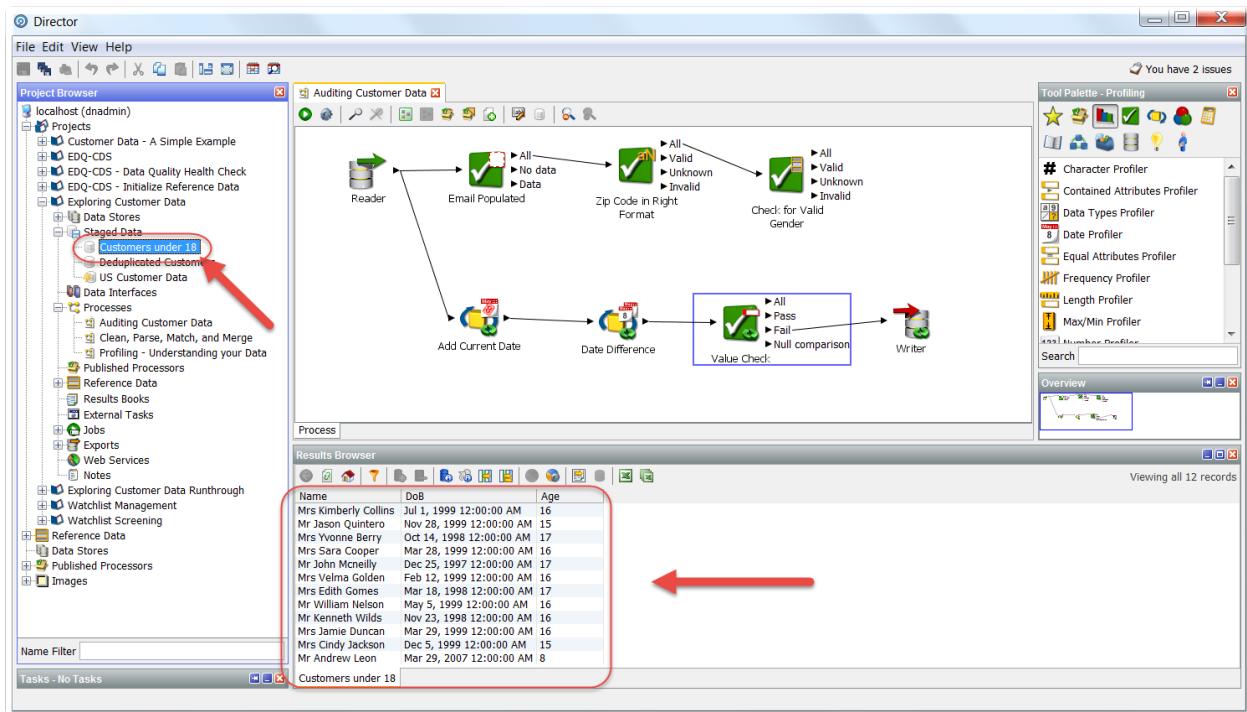


47. Sau đó, bạn sẽ nhận thấy **Tên**, **Ngày sinh**, và **Tuổi** là các thuộc tính được ánh xạ tới **Dữ liệu được phân loại** thiết lập bạn vừa tạo. Nhấp vào **ĐƯỢC RỒI** để hoàn tất việc cấu hình Writer



48. Nhấp chuột ở góc trên bên trái của Project Canvas để chạy quy trình

Bạn sẽ nhận thấy sự tập trung của Giám đốc chuyển trả lại **Trình duyệt dự án** và Dữ liệu được phân
giai đoạn mới của bạn, **Khách hàng dưới 18 tuổi** được chọn. **Trình duyệt kết quả** cũng đã thay đổi để
hiển thị **6 Khách hàng** người không vượt qua được Kiểm tra giá trị lớn hơn hoặc bằng 18.

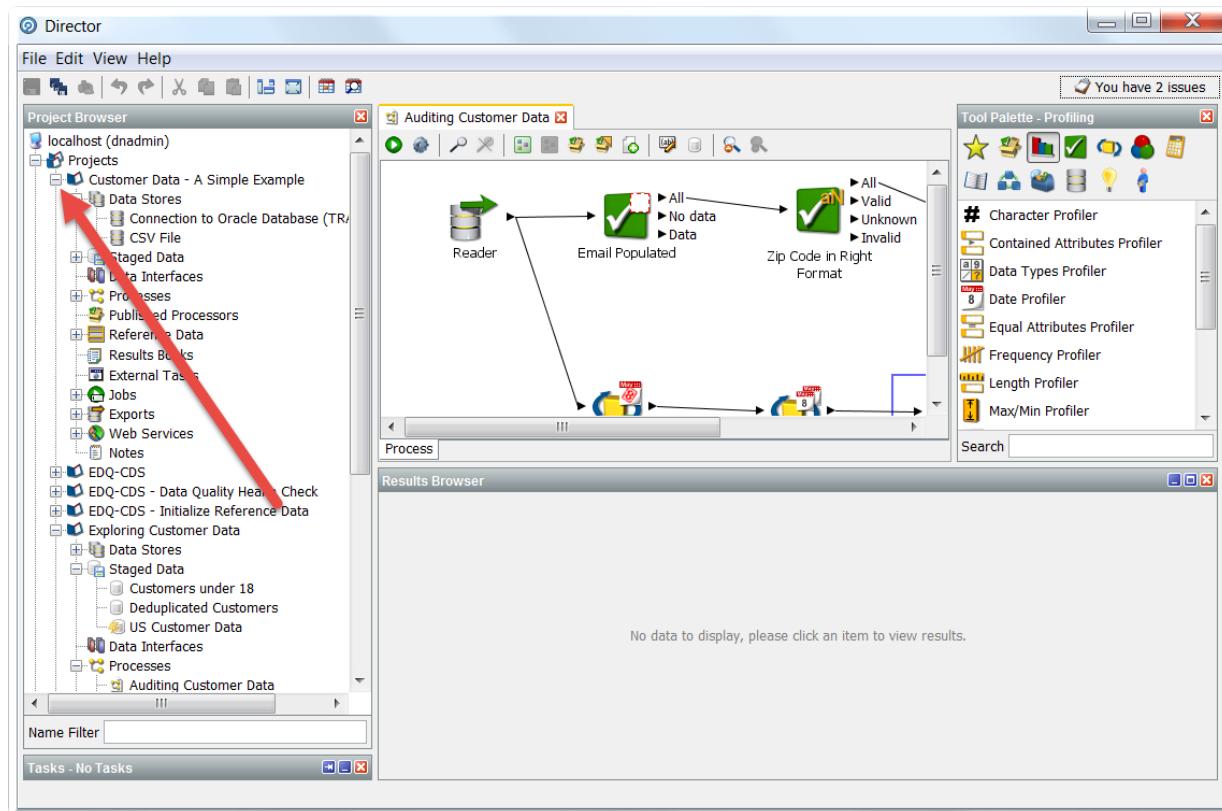


Trong Lab tiếp theo, chúng ta sẽ có thể sử dụng những gì đã học về dữ liệu của mình trong các giai đoạn Profile và Auditing để Clean, Enrich, Parse, Transform, Match và Merge thông tin. Trước đó, chúng ta cần chuẩn bị thêm hai bộ Reference Data để có thể chuẩn hóa vấn đề tìm thấy với cột Country.

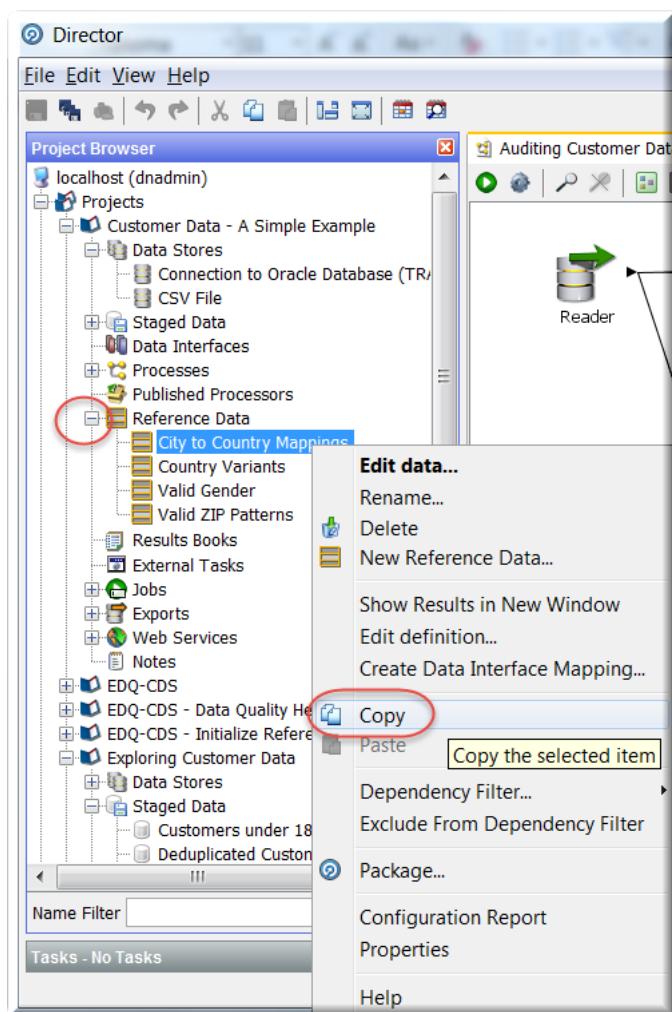
Chúng ta đã xem xét việc tạo ra các vấn đề để tạo ra một môi trường cộng tác và chúng ta sẽ đi sâu hơn vào vấn đề đó trong Phòng thí nghiệm 5. Ngoài ra còn có những cách khác để biến EDQ thành một công cụ cộng tác. Ví dụ, Dữ liệu tham chiếu và các mục khác trong Trình duyệt dự án có thể dễ dàng được chia sẻ bằng cách chỉ cần sao chép và dán từ dự án này sang dự án khác.

Để tiết kiệm thời gian, chúng tôi sẽ sử dụng lại Dữ liệu tham chiếu được tạo trong một dự án khác.

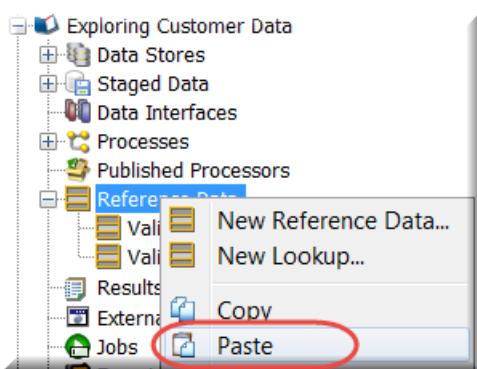
49. Điều hướng đến Trình duyệt Dự án và tìm Dự án có tiêu đề **Dữ liệu khách hàng - Một ví dụ đơn giản**



50. Mở rộng **Dữ liệu tham khảo** mục và nhấp chuột phải để **Sao chép** **cái Bản đồ thành phố đến quốc gia**



51. Quay trở lại **Dự án - Khám phá dữ liệu khách hàng** và tìm thấy **Dữ liệu tham khảo** mục, nhấp chuột phải và **Dán** nội dung của clipboard



52. Quay lại **Trình duyệt dự án** và tìm Dự án có tiêu đề **Dữ liệu khách hàng - Một ví dụ đơn giản**

53. Mở rộng **Dữ liệu tham khảo** mục nếu nó chưa có và nhấp chuột phải và **Sao chép** cái **Biển thể quốc gia** Dữ liệu tham khảo

54. Quay trở lại **Dự án - Khám phá dữ liệu khách hàng** và nhấp chuột phải vào **Dữ liệu tham khảo** và **Dán** nội dung của clipboard

55. Trong **Dữ liệu tham khảo** phần của Dự án của bạn, **Khám phá dữ liệu khách hàng**, nhấp vào **Bản đồ thành phố đến quốc gia**

Lưu ý trong **Trình duyệt kết quả** đây là một bộ dữ liệu tham chiếu hai cột. **Thành phố** (màu vàng) biểu thị Cột tra cứu, tương tự như hai bộ dữ liệu tham chiếu mà chúng tôi đã tạo ở các bước trước cho Giới tính và Mã bưu chính. **Quốc gia** (cột màu xanh lá cây) biểu thị Cột trả về, mà chúng ta vẫn chưa khám phá. Trong trường hợp này, bất cứ nơi nào giá trị Cột tra cứu chứa dữ liệu, **Bản đồ thành phố đến quốc gia** Bộ dữ liệu tham chiếu có thể được sử dụng để trả về Quốc gia của Thành phố đó. Điều này sẽ giúp chúng tôi chuẩn hóa các giá trị tìm thấy trong cột Quốc gia ở các bước tiếp theo.

City	Country
Tokyo	Japan
Seoul	Korea (South)
Mexico City	Mexico
Delhi	India
Bombay	India
New York	United States of America
Sao Paulo	Brazil
Manila	Philippines
Los Angeles	United States of America
Shanghai	China
Osaka	Japan
Calcutta	India

56. Nhấp vào **Biển thể quốc gia** Dữ liệu tham chiếu trong Dự án, **Khám phá dữ liệu khách hàng** trong Trình duyệt Dự án

Lưu ý rằng đây cũng là một tập dữ liệu tham chiếu hai cột với **Các biến thể quốc gia (Hoa Kỳ, US, USA, v.v.)** như Cột tra cứu và một **Giá trị chuẩn hóa (Hoa Kỳ)** như Cột trả về. Nếu bạn lật lại trang mà chúng ta đã tạo Biểu đồ tròn và Biểu đồ thanh cho bài tập Phân tích quốc gia, bạn sẽ nhớ rằng chúng ta có một số biểu diễn của cùng một quốc gia Logic trong tập dữ liệu. Dữ liệu tham chiếu này sẽ cho phép chúng ta chuẩn hóa nhiều biểu diễn của một quốc gia duy nhất thành một biểu diễn chuẩn hóa duy nhất của quốc gia – tăng đáng kể 'mức độ phù hợp khi sử dụng' cho dữ liệu khi thực hiện cho tất cả các cột có liên quan.

Người hưởng lợi chính: Tác giả và người tiêu dùng bảng điều khiển phân tích 'Không của bảng điều khiển Phân tích. chỉ là phân tích, mà là phân tích chính xác'.



Đây là bản sao duy nhất của Dữ liệu tham chiếu. Nếu Dữ liệu tham chiếu được sửa đổi trong dự án của bạn, nó sẽ vẫn nguyên vẹn cho dự án khác. Nếu bạn tình cờ tìm thấy một biến thể bổ sung của một quốc gia trong Trình duyệt kết quả, hãy nhấp chuột phải vào Dữ liệu tham chiếu để chỉnh sửa và thêm các hàng cần thiết. Điều này cho phép bạn sửa đổi quy trình và/hoặc bộ xử lý sử dụng Dữ liệu tham chiếu mà không cần thực hiện các thay đổi lớn có thể tốn thời gian.

Phòng thí nghiệm 3: Tạo và tự động hóa các giải pháp chuẩn hóa và cải tiến dữ liệu

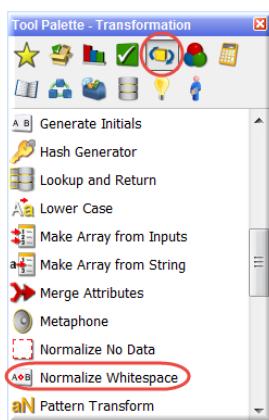
Trong phòng thí nghiệm này, chúng ta sẽ thực hiện những điều sau bằng cách sử dụng Bộ xử lý EDQ:

- Chuẩn hóa
- Biến đổi
- Phân tích
- Làm giàu

Tất cả các bước này đều 'quan trọng để thành công' cho sáng kiến lấy dữ liệu làm trung tâm bằng cách tạo ra "Dữ liệu phù hợp để sử dụng" và "Phân tích chính xác".

Tạo quy trình chuẩn hóa

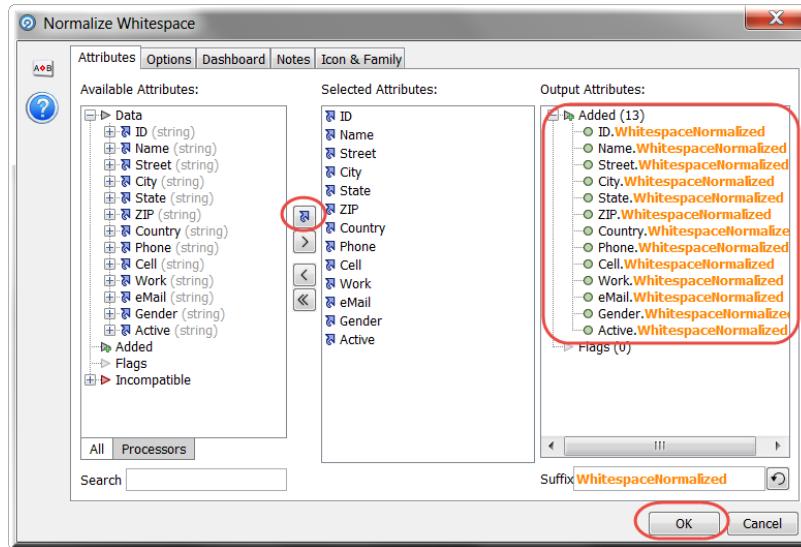
1. Tạo một quy trình mới trong **Dự án (Khám phá dữ liệu khách hàng)** trong **Trình duyệt dự án** bằng cách nhấp chuột phải vào **Các quy trình** và nhấp chuột **Quy trình mới...**
2. Chọn **Dữ liệu khách hàng Hoa Kỳ** và nhấp vào **Tiếp theo >**. Nhấp chuột **Tiếp theo >** không cần thêm bất kỳ Hồ sơ nào
3. Đặt tên cho quy trình của bạn **Làm sạch, Phân tích, So khớp và Hợp nhất** sau đó nhấp vào **Hoàn thành**
4. Quay lại **Bảng công cụ** ở phía bên phải của **Giám đốc** và tìm thấy **Chuẩn hóa khoảng trắng** bộ xử lý bằng cách nhấp vào **Sự biến đổi**



5. Kết nối **Người đọc** đến **Chuẩn hóa khoảng trắng** Bộ xử lý

6. Các **Hộp thoại Chuẩn hóa Khoảng trắng** xuất hiện. Nhấp vào **Chọn tất cả** nhấp chuột **ĐƯỢC RỒI**. Tiếp theo nhấp vào **Chạy** biểu tượng trên thanh công cụ để bắt đầu quá trình

 biểu tượng và

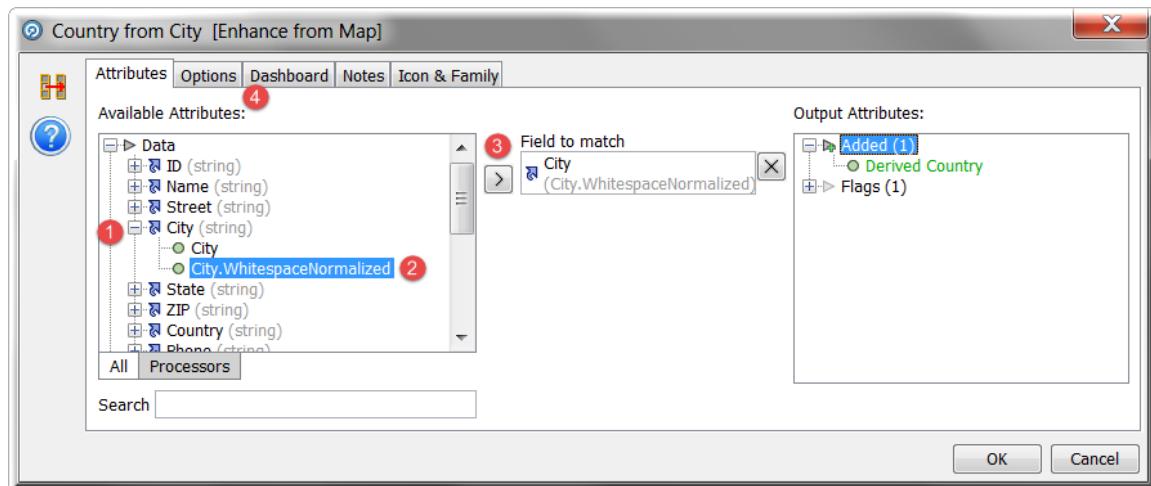


Tiếp theo chúng ta sẽ bắt đầu quá trình Chuẩn hóa cột Quốc gia.

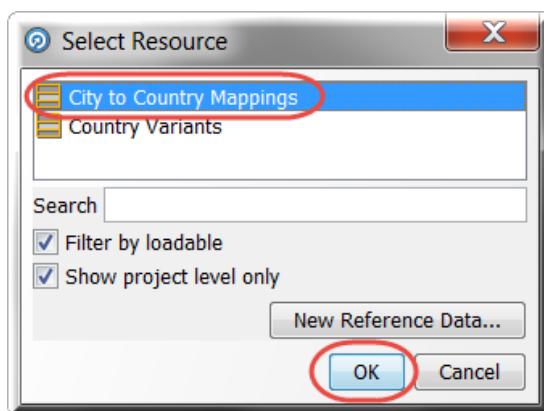
7. Tìm **Cải thiện từ Bản đồ** Bộ xử lý từ **Bảng công cụ** bằng cách sử dụng **Tìm kiếm** trường văn bản: loại **Cải thiện từ Bản đồ**. Kéo và thả Bộ xử lý vào khung vẽ Quy trình. Nhấp chuột phải vào **Cải thiện từ Bản đồ** và chọn **Đổi tên** để đổi tên thành **Đất nước từ thành phố**

8. Kết nối tam giác cuối từ **Chuẩn hóa khoảng trắng** Bộ xử lý cho **Đất nước từ thành phố** Bộ xử lý

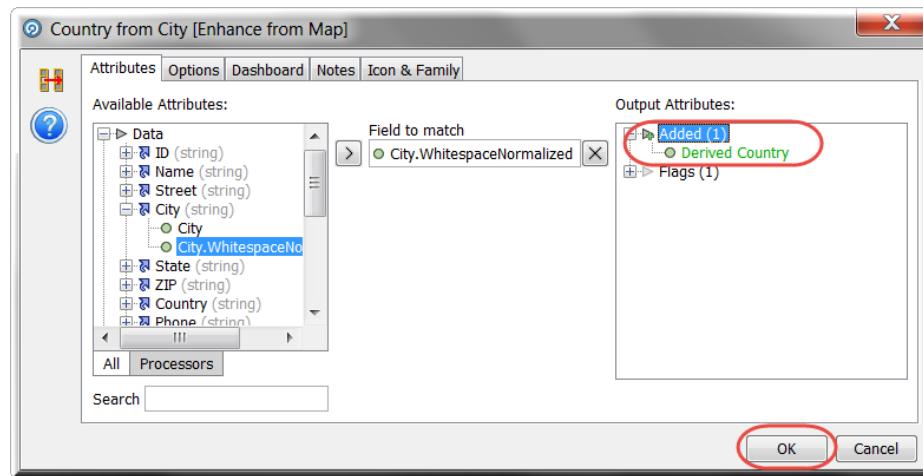
9. Trong **Hộp thoại Quốc gia từ Thành phố**, mở rộng **Thành phố** lĩnh vực từ **Thuộc tính có sẵn** để quan sát các giá trị siêu dữ liệu bổ sung cho lựa chọn được tạo bởi Bộ xử lý khoảng trắng chuẩn hóa. Nhấp đúp vào **Thành phố** để thêm **Thành phố** đến **Trường để phù hợp**. Tiếp theo, nhấp vào **Tùy chọn** tab ở đầu hộp thoại



10. Nhấp vào Browse mà chúng ta *** nút bên trong **Bản đồ giá trị** khu vực để kết nối Dữ liệu tham chiếu Chọn **Bản đồ thành phố đến quốc gia**, nhấp chuột **ĐƯỢC RỒI** để tiếp tục



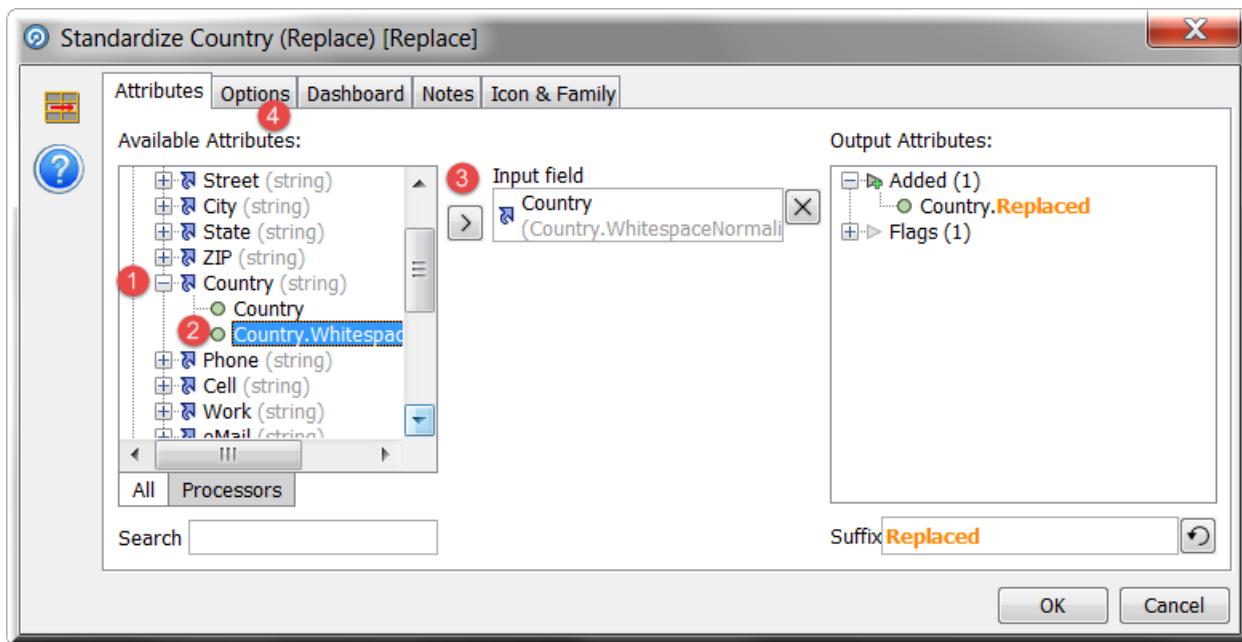
11. Nhấp chuột **Thuộc tính** ở góc bên trái của hộp thoại để điều chỉnh tên của **Thuộc tính đầu ra: EnhancedResult** bằng cách nhấp đúp vào **Quốc gia có nguồn gốc**, sau đó nhấp vào **ĐƯỢC RỒI**



12. Nhấp chuột để bắt đầu **Đất nước từ thành phố** quá trình

Bạn sẽ thấy một số kết quả chưa được cải tiến, nhưng chúng ta vẫn chưa xong! Hãy thoải mái xem xét kỹ hơn các giá trị Cải tiến và Chưa cải tiến để xem bộ xử lý này đã làm gì.

13. Quay trở lại **Bảng công cụ** để tìm kiếm **Thay thế**. Kéo và thả **Thay thế** Bộ xử lý cho Project Canvas. Nhấp chuột phải vào **Thay thế** Bộ xử lý và chọn **Đổi tên** để đổi tên nó **Chuẩn hóa quốc gia**
14. Các **Chuẩn hóa quốc gia** hộp thoại cấu hình xuất hiện. Kết nối **Tất cả** kết thúc tam giác từ **Đất nước từ thành phố** Quá trình. Nhấp đúp vào **Quốc gia** trường và thêm **Quốc gia**. **Khoảng trăng Chuẩn hóa** vào Trường nhập liệu. Sau đó nhấp vào **Tùy chọn** tab ở đầu hộp thoại để thiết lập Thay thế



15. Nhập vào Browse Reference Data nút bên cạnh **Thay thế** để thêm vào **Biến thể quốc gia**
nhập vào **ĐƯỢC RỒI**



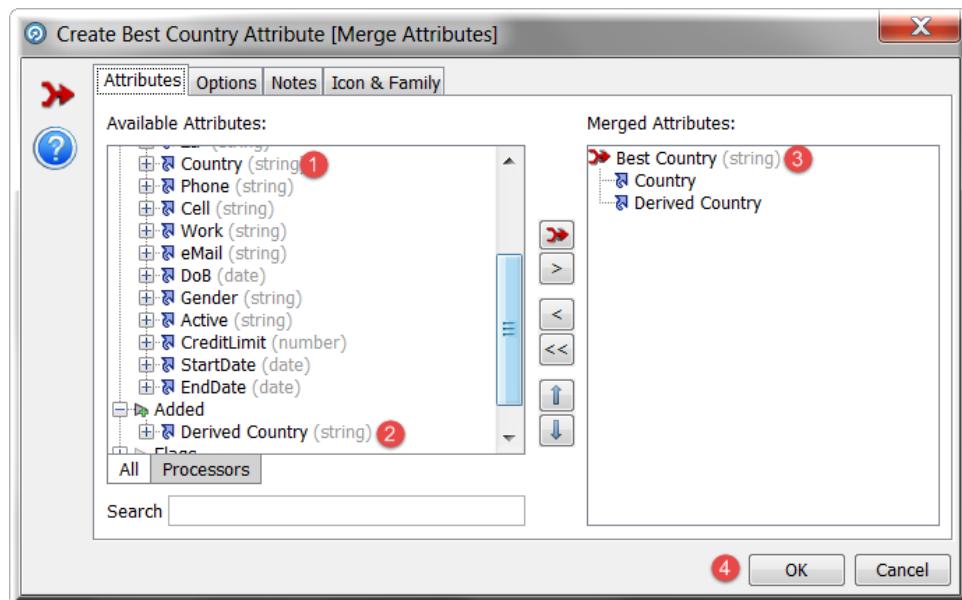
16. Nhấp vào **Chạy** nút trên thanh công cụ để bắt đầu quá trình

17. Quay trở lại **Bảng công cụ** và tìm kiếm **Hợp nhất**. Kéo và thả **Hợp nhất các thuộc tính** Bộ xử lý vào Project Canvas và đổi tên nó thành **Tạo Thuộc tính Quốc gia Tốt nhất**

Chúng tôi sẽ sử dụng điều này để kết hợp các cột Quốc gia và Quốc gia phái sinh thành một thuộc tính duy nhất có thể được sử dụng ở giai đoạn sau của Quy trình, do đó, tiến gần hơn đến 'nguồn thông tin xác thực duy nhất'.

18. Kết nối **Tất cả** kết thúc tam giác từ **Chuẩn hóa quốc gia** đến **Tạo Thuộc tính Quốc gia Tốt nhất** Bộ xử lý.

Nhấp đúp vào **Quốc gia** và **Quốc gia có nguồn gốc** từ các Thuộc tính Có sẵn để thêm vào **Thuộc tính được hợp nhất**. Cuối cùng, đổi tên Thuộc tính đã hợp nhất thành **Đất nước tốt nhất**, sau đó nhấp vào **ĐƯỢC RỜI** để tiếp tục

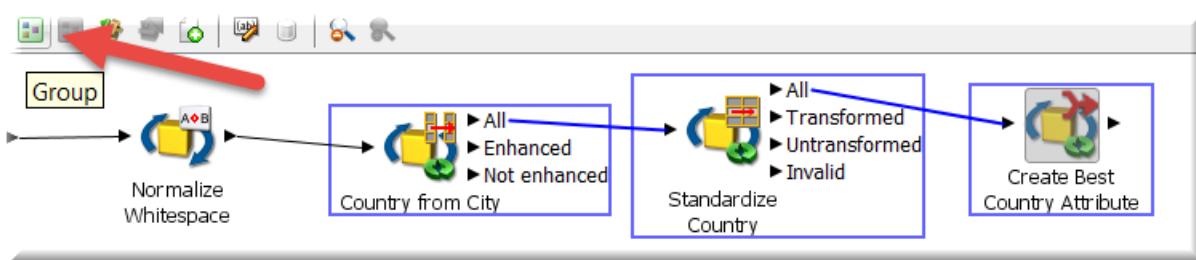


19. Nhấp vào **Chạy** nút ở góc trên bên trái của Project Canvas và sau đó nhấp vào **Tạo Thuộc tính Quốc gia Tốt nhất** Bộ xử lý để xem kết quả trong **Trình duyệt kết quả**

Results Browser		
Job:	Clean, Parse, Match, and Merge	
Country.Replaced	Derived Country	Best Country
United States of America	United States of America	United States of America
United States of America	United States of America	United States of America
United States of America	United States of America	United States of America
United States of America	United States of America	United States of America
United States of America	United States of America	United States of America
United States of America	United States of America	United States of America
United States of America	United States of America	United States of America
United States of America	United States of America	United States of America
United States of America	United States of America	United States of America

Lưu ý cách hai cột, **Quốc gia**. **Đã thay thế** và **Quốc gia có nguồn gốc** hiện đã được hợp nhất thành một cột duy nhất, **Đất nước tốt nhất**.

20. Bạn cũng có thể nhóm nhiều bộ xử lý lại với nhau bằng cách tô sáng các bộ xử lý mong muốn để nhóm và nhấp vào **Nút nhóm** trong Thanh công cụ, nhấp và kéo một hộp xung quanh **Đất nước từ thành phố**, **Chuẩn hóa quốc gia**, và **Tạo Thuộc tính Quốc gia Tốt nhất** bộ xử lý. Sau đó nhấp vào **Nút nhóm**  trong thanh công cụ



21. Sau đó, bạn có thể nhấp vào **Nhóm** trong hộp màu xanh lá cây nó hình thành để đổi tên nhóm. Trong trường hợp này, đổi tên nó **Chuẩn hóa/Thêm quốc gia**

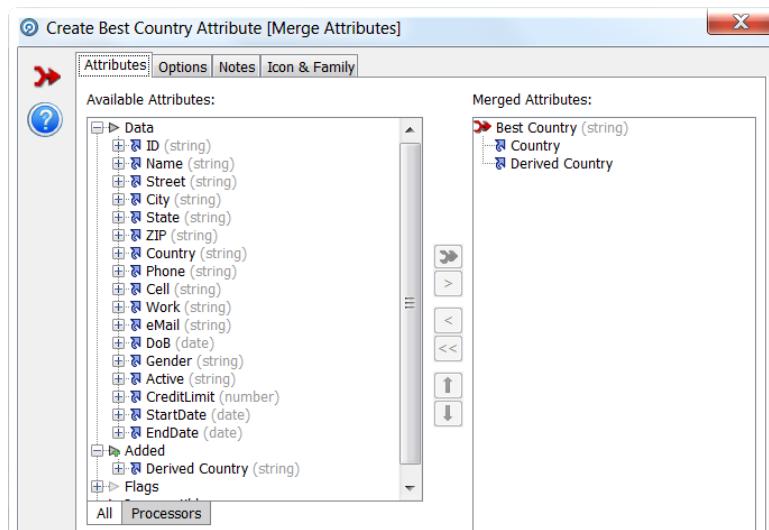


Bộ xử lý nhóm có thể được xuất bản dưới dạng Bộ xử lý để những người dùng EDQ khác có thể sử dụng Bộ xử lý trong các dự án của họ. Chúng tôi sẽ không đề cập đến khả năng này trong hướng dẫn phòng thí nghiệm này.

Trước khi tiếp tục, chúng ta hãy thêm một Frequency Profiler vào cuối **Chuẩn hóa/Thêm quốc gia** nhóm để kiểm tra tiến độ làm cho bộ dữ liệu này phù hợp để sử dụng (bằng cách chuẩn hóa dữ liệu)

22. Quay trở lại **Bảng công cụ** và nhập **Tần số Profiler** trong **Tìm kiếm** trường văn bản. Kéo và thả **Tần số Profiler** vào khung Project và kết nối hình tam giác cuối từ **Tạo Thuộc tính Quốc gia Tốt nhất** đến **Tần số Profiler** và hộp thoại cấu hình xuất hiện

23. Nhấp đúp vào **Đất nước tốt nhất** thuộc tính để thêm nó vào **Thuộc tính được chọn** cột và nhấp vào **ĐƯỢC RỒI**.



24. Đổi tên Frequency Profiler bằng cách nhấp chuột phải, chọn **Đổi tên** và nhập **Tần suất Hồ sơ Quốc gia**. Nhấp vào **Chạy** nút trên thanh công cụ để bắt đầu quá trình

Results Browser		
Value	Count	%
United States of America	4858	89.3%
	520	9.6%
Canada	50	0.9%
Italy	3	<0.1%
Ireland	2	<0.1%
China	1	<0.1%
United Kingdom	1	<0.1%
Ukraine	1	<0.1%
Greece	1	<0.1%
Egypt	1	<0.1%



Quan sát các giá trị thống nhất cho các quốc gia. Chúng tôi đã chuẩn hóa năm (5) cách biểu diễn khác nhau của Hoa Kỳ (USA, US, USA, United States và US) thành một và chỉ một cách biểu diễn Chuẩn hóa: **Hoa Kỳ**. Nếu không có sự chuẩn hóa, một trong hai điều sau chắc chắn sẽ xảy ra:

- Đếm thiếu doanh số tại Hoa Kỳ khi báo cáo trong Dashboard Analytics. Tác giả Dashboard sẽ chọn một trong các giá trị Quốc gia. Nhưng còn giá trị của Hoa Kỳ và Hoa Kỳ thì sao? Doanh số đó sẽ được báo cáo ở Quốc gia nào?

- Tác giả của Dashboard phải biết trước tất cả các giá trị không chuẩn hóa của các Quốc gia hiện có và tự xử lý tất cả các bất thường. Khả năng điều đó xảy ra đối với tất cả các bất thường trong từng Thuộc tính trong dữ liệu là bao nhiêu? Ngay cả khi có xảy ra - thì phải mất bao lâu nữa để phát triển và thử nghiệm các Dashboard này so với các Dashboard được viết dựa trên dữ liệu Chuẩn hóa? Hãy nghĩ đến năng suất! Hãy nghĩ đến giá trị kinh doanh của kết quả Dashboard không chính xác?

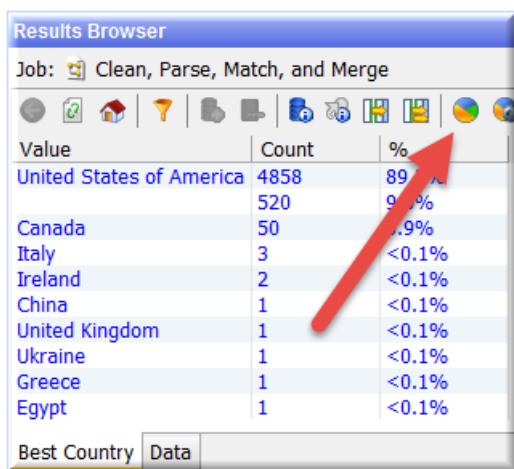


Ngoài ra, lưu ý rằng vẫn còn 520 hoặc 9,6% giá trị quốc gia vẫn cần được quan tâm.

25. Nhấp chuột phải vào **520** và **Tạo văn đề...**

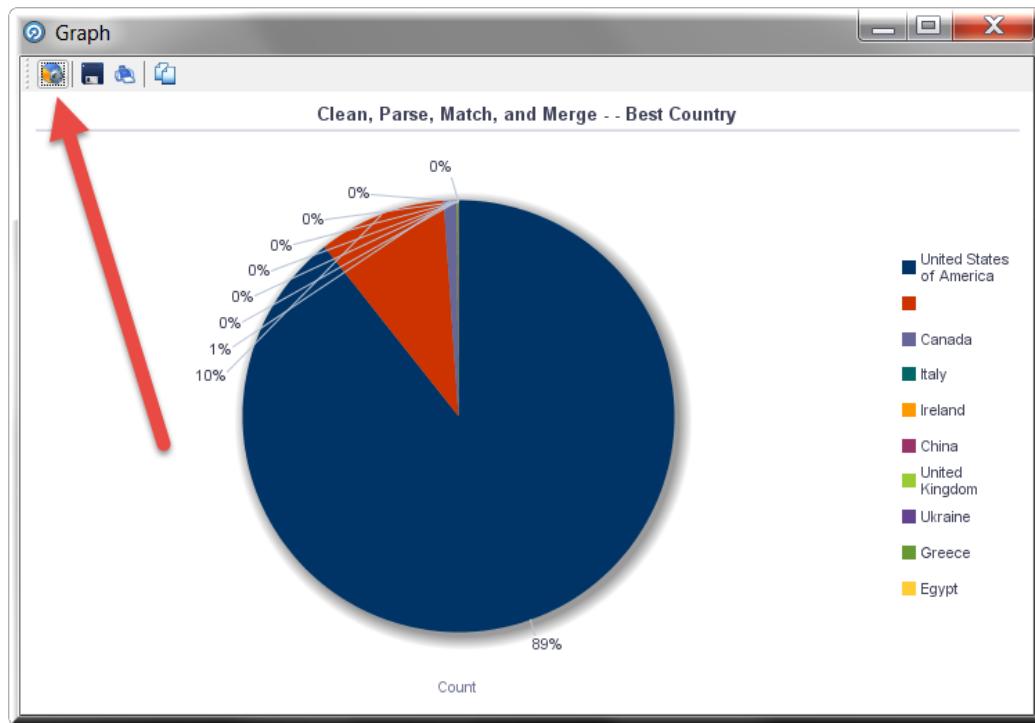
26. Nhập mô tả vào **Mô tả văn đề** Khu vực văn bản cho vấn đề này để chúng ta có thể quay lại trong bài tập Lab tiếp theo – **520 mục văn tồn tại sau quá trình chuẩn hóa/thêm quốc gia**, chọn người dùng cho **Được giao cho** và nhấp vào **ĐƯỢC RỒI** để đóng hộp thoại vấn đề

27. Nhấp vào **Kết quả biểu đồ** biểu tượng trên thanh công cụ của **Trình duyệt kết quả**

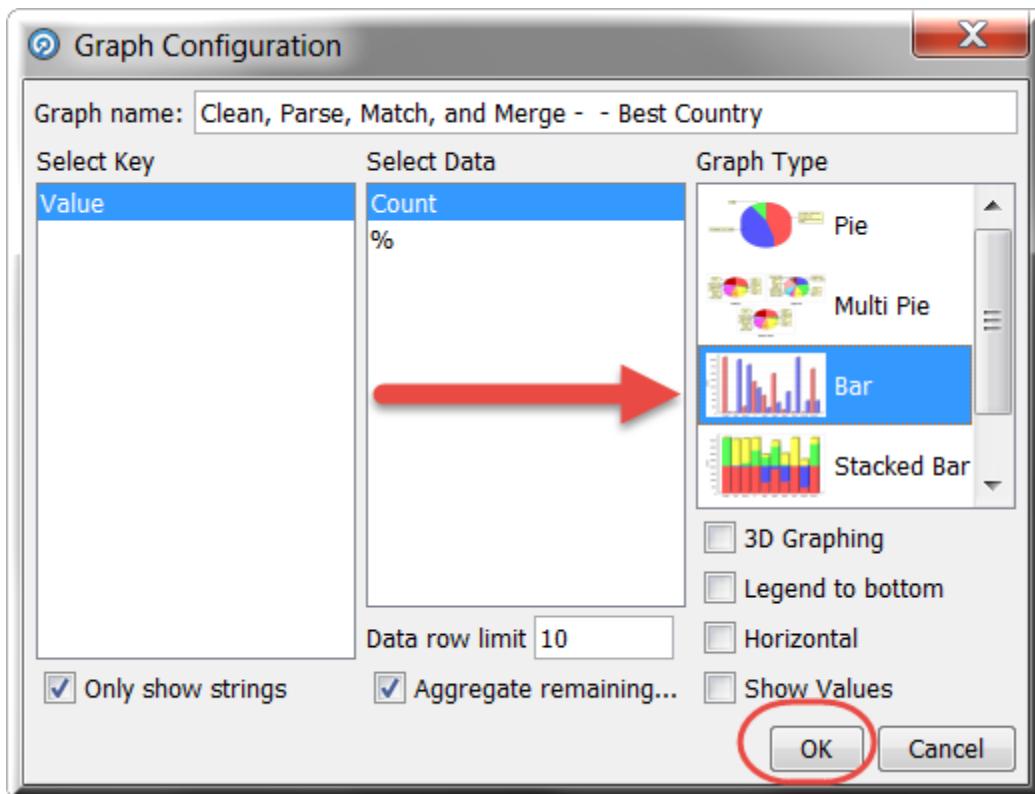


Lưu ý biểu đồ này trông như thế nào khác biệt đáng kể từ cái chúng tôi tạo ra trong Lab 1b: Hồ sơ. Các đại diện khác nhau của Hoa Kỳ hiện đã được chuẩn hóa thành Hoa Kỳ

28. Nhấp vào **Nút Cấu hình đồ thị** trên thanh công cụ của biểu đồ

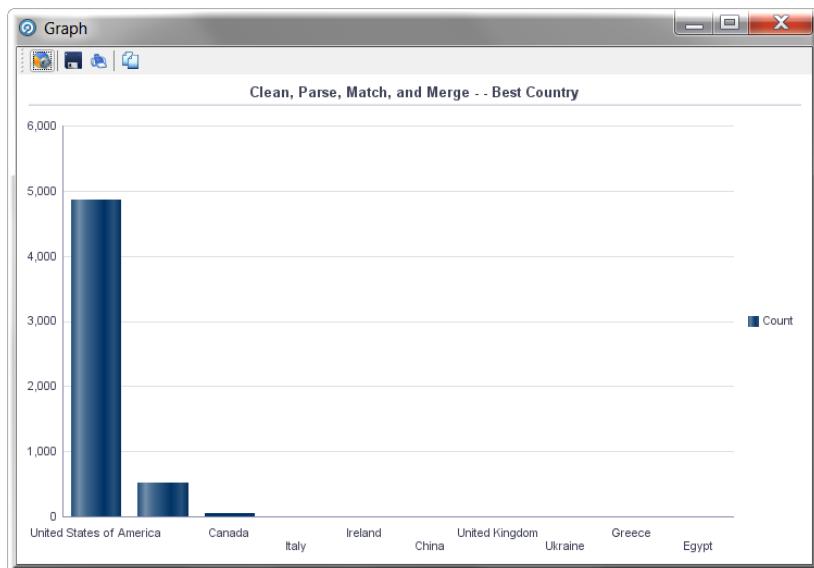


29. Thay đổi **Kiểu đồ thị** ĐẾN **Thanh**, nhấp chuột **ĐƯỢC RỒI** để tiếp tục

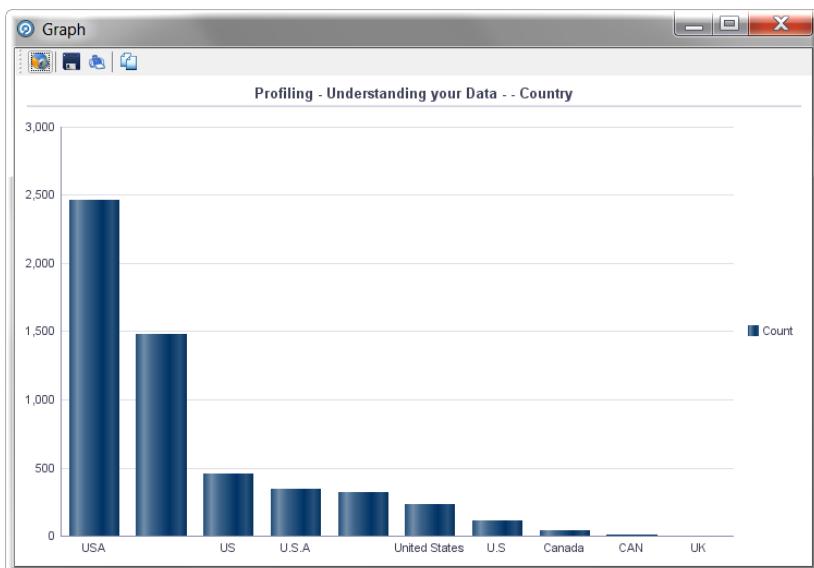


Hãy quan sát sự khác biệt của biểu đồ sau đây so với biểu đồ được tạo trước đó trước khi Chuẩn hóa.
Chúng tôi hoàn toàn có thể nói rằng cột Quốc gia có khả năng sử dụng cao hơn nhiều cho các sáng kiến dữ liệu hạ nguồn tiêu thụ và/hoặc báo cáo về dữ liệu. Ngoài ra, còn có những quốc gia khác mà chúng tôi chưa thấy trước đây cũng đã được thêm vào.

Sau khi chuẩn hóa



Trước khi chuẩn hóa

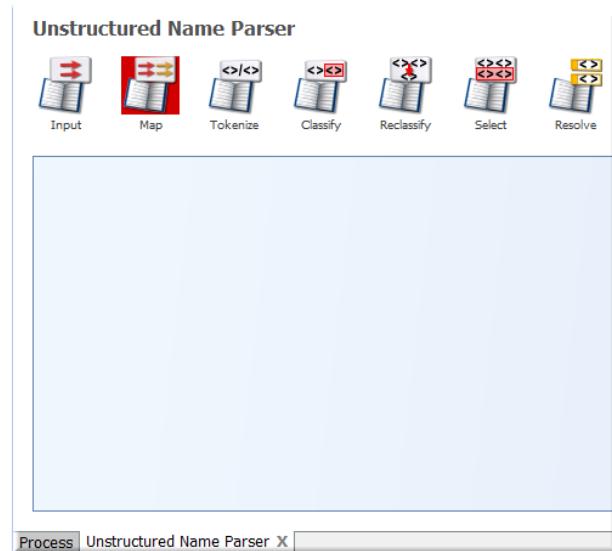


Tiếp theo chúng ta sẽ bắt đầu kiểm tra và áp dụng cấu trúc cho trường Tên.

30. Điều hướng đến **Bảng công cụ** và tìm kiếm bằng cách nhập **Trình phân tích tên không có cấu trúc**. Kéo và thả **Trình phân tích tên không có cấu trúc** vào Project Canvas và kết nối **Chuẩn hóa khoảng trắng** điểm cuối đến **Trình phân tích tên không có cấu trúc**
31. Các **Hộp thoại Phân tích tên không có cấu trúc** xuất hiện, chọn **Tên** từ **Thuộc tính có sẵn** và thêm nó vào **Thuộc tính được chọn**, nhấp chuột **ĐƯỢC RỒI** để tiếp tục

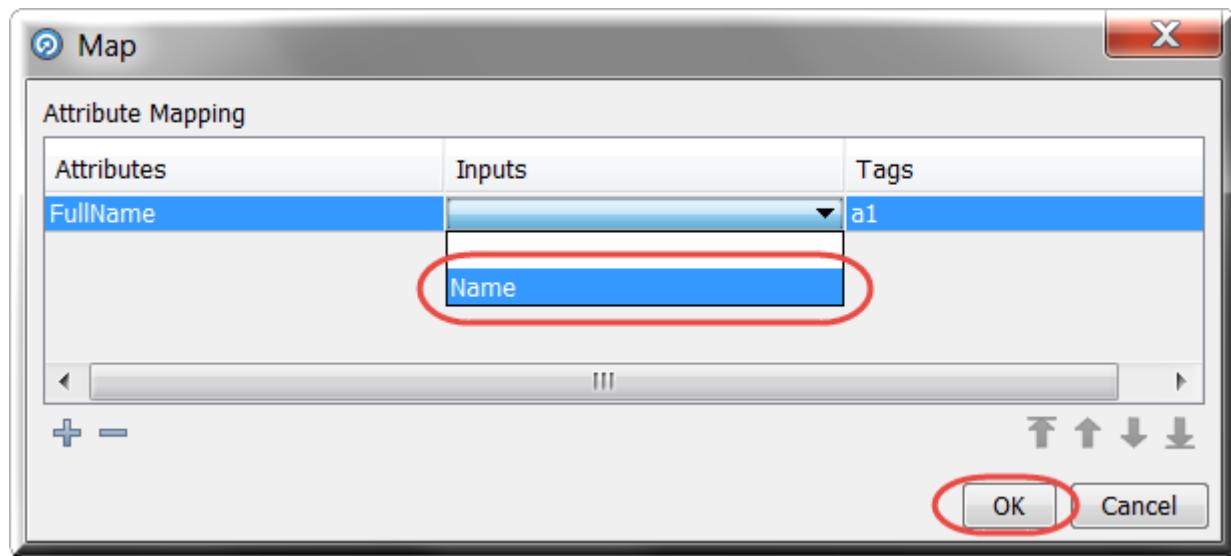


Bộ xử lý Parse bao gồm một số màn hình cấu hình. Bộ xử lý này được cấu hình sẵn để làm việc với Names, vì vậy chúng ta sẽ không phải lo lắng về việc cấu hình từng bộ xử lý phụ. Ngoài ra, bạn sẽ thấy Project Canvas có nhiều tab ở góc dưới bên trái (phía trên **Trình duyệt kết quả**). Bạn có thể chuyển đổi qua lại giữa Project Canvas và **Trình phân tích tên không có cấu trúc** cấu hình.



Ở đầu Project Canvas, bạn sẽ thấy một số bộ xử lý phụ: **Đầu vào**, **Bản đồ**, **Phân loại**, **Phân loại lại**, **Chọn và Giải quyết**. Bất kỳ bộ xử lý phụ nào là **được tô bóng màu đỏ** cần phải được cấu hình.

32. Nhấp đúp vào **Bản đồ** để cấu hình nó. Vì chúng tôi đã chọn một thuộc tính cho bộ xử lý này, nên nó chỉ có một đầu vào duy nhất – **Họ và tên đầy đủ**. Chọn **Tên** thuộc tính dưới **Đầu vào** thả xuống và nhấp vào **ĐƯỢC RỒI** để tiếp tục



33. Nhấp vào **Chạy** nút trên thanh công cụ để chạy quy trình. Sau một vài phút, bộ xử lý sẽ hiển thị các mẫu mã thông báo đã chọn được tìm thấy trong trường Tên trong **Lựa chọn tab** trong Trình duyệt kết quả

Results Browser				
FullName	Exact Rule	Fuzzy Rule	Count	%
<valid Title>_<valid Given>_<valid Family>	3	5287	97.2	
<Valid Title> '_<valid Given>_<valid Family>	3	73	1.3	
<valid Given>_<valid Family>	6	38	0.7	
<valid Title>_<possible Given>_<valid Family>	3	11	0.2	
<valid Title>_<possible Given>_<A>_<valid And>_<valid...	50	4	<0.1	
<valid Entity>	13	4	<0.1	
<valid Title>_<valid Initial>_<valid Family>	11	3	<0.1	
<valid Title>_<possible Given>_<valid Initial>_<valid Fa...	9	3	<0.1	
<valid Title>_<possible Given>_<A><S><A>		2	<0.1	
<valid Title>_<valid Initial>_<valid Initial>_<valid Family>	10	2	<0.1	
<valid Title>_<valid Given>_<valid Middle>_<valid Family>	1	2	<0.1	
<possible Given>_<A>_<valid And>_<valid Given>_<vali...	54	2	<0.1	
<valid Title>_<valid Title>_<valid Family>	28	2	<0.1	
<valid Title>_<valid And>_<valid Title>_<A>_<A>	48	1	<0.1	
<valid Title>_<valid Family>	25	1	<0.1	
<valid Title>_<valid Given>_<possible Middle>_<valid Fa...	1	1	<0.1	
Ambiguity (2)		1	<0.1	
<valid Initial>_<valid And>_<valid Initial>_<valid Family>	19	1	<0.1	

Lưu ý rằng hơn 97% bản ghi có mẫu <Tiêu đề hợp lệ>_<Tên hợp lệ>_<Họ hợp lệ>.

34. Khoan sâu vào mẫu bằng cách nhấp vào <Tiêu đề hợp lệ>_<Đã cho hợp lệ>_<Họ hợp lệ> để xem trình phân tích cú pháp này chia tách như thế nào Tên thuộc tính

Results Browser

Job: Clean, Parse, Match, and Merge

Name.WhitespaceNormalized	UnclassifiedData	P1_Prefix	P1_First	P1_Middle	P1_Last
Mrs Christine Hunt		Mrs	Christine		Hunt
Mr Walter Terrell		Mr	Walter		Terrell
Mr Lester Young		Mr	Lester		Young
Mr John Mcalister		Mr	John		Mcalister
Mrs Phyllis Ford		Mrs	Phyllis		Ford
Mrs Marilyn Tan		Mrs	Marilyn		Tan
Mrs Mary Hill		Mrs	Mary		Hill
Mrs Tasha Trower		Mrs	Tasha		Trower
Mr Brian Robles		Mr	Brian		Robles
Mr Richard Brown		Mr	Richard		Brown
Mr Perry Seibert		Mr	Perry		Seibert
Mrs Anna Sanroman		Mrs	Anna		Sanroman
Mrs Edith Reed		Mrs	Edith		Reed
Mr Darren Hamilton		Mr	Darren		Hamilton
Mr Nathaniel Bakken		Mr	Nathaniel		Bakken
Mr John Morris		Mr	John		Morris

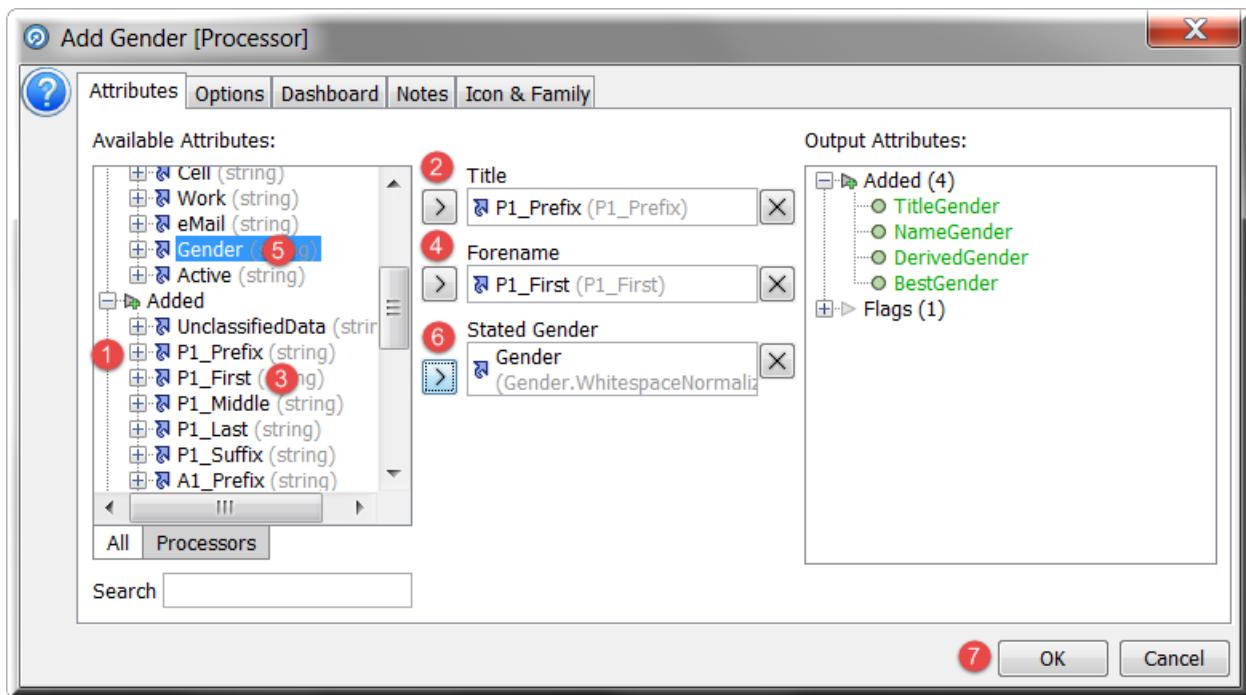
Bây giờ chúng ta có thể sử dụng các thuộc tính bổ sung này (P1_Prefix, P1_First, v.v.) ở phần sau của quy trình này. Ví dụ, chúng ta có thể lấy được Gender dựa trên cột P1_Prefix để cải thiện thêm chất lượng dữ liệu Gender như đã lưu ý trong Lab 1.

35. Nhấp vào X trên **Trình phân tích tên không có cấu trúc** ở góc dưới bên trái của Project Canvas để quay lại **Quy trình Canvas**

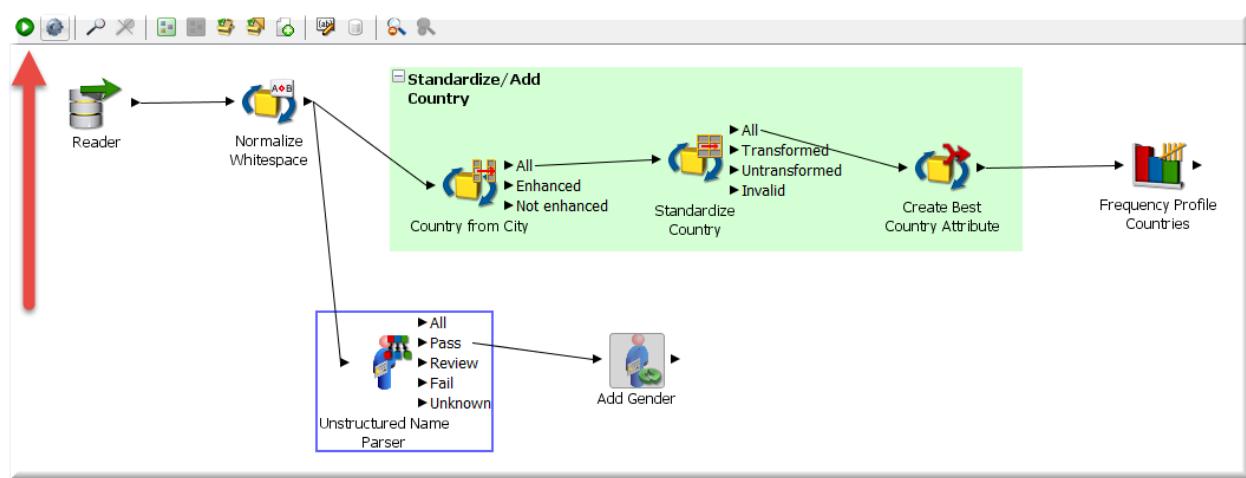


36. Quay trở lại **Bảng công cụ** và tìm kiếm bằng cách nhập **Thêm Giới tính**. Kéo và thả **Thêm Giới tính** bộ xử lý đến Project Canvas và kết nối **Vượt qua** đầu ra tam giác từ **Trình phân tích tên không có cấu trúc** vào đầu vào của **Thêm Giới tính** bộ xử lý

37. Chọn **P1_Tiền tố** và nhấn > biểu tượng để di chuyển nó đến **Tiêu đề**. Lựa chọn **P1 Đầu tiên** và di chuyển đến **Tên**. Lựa chọn **Giới tính** và di chuyển nó đến **Giới tính đã nêu**, sau đó nhấp vào **ĐƯỢC RỒI** để tiếp tục

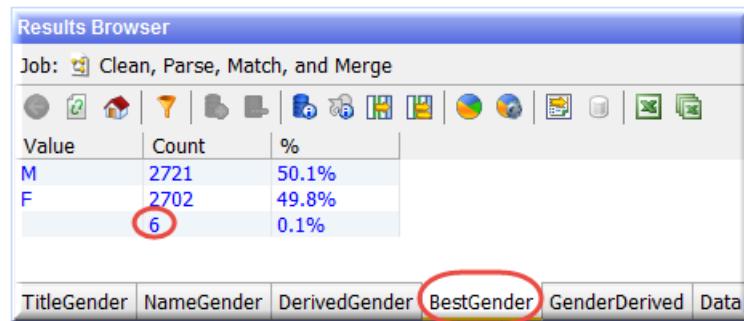


38. Nhấp vào **Chạy** biểu tượng trên thanh công cụ để bắt đầu quá trình

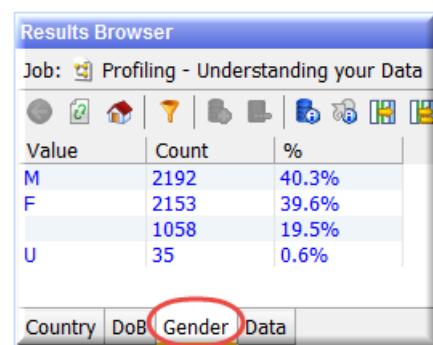


39. Trong **Trình duyệt kết quả**, nhấp vào **Giới tính tốt nhất** tab. Điều này hiển thị số lượng và tỷ lệ thành công và không thành công của các bản ghi có nguồn gốc từ Tiêu đề và tên. Nhấp vào **Đếm** cột để phân tích sâu hơn

Sau khi thêm giới tính (giới tính được làm giàu/chuẩn hóa):



Trước khi thêm giới tính (Ảnh chụp màn hình từ Phòng thí nghiệm 1b)...Dữ liệu không phù hợp để sử dụng



i Chúng tôi đã tự động làm giàu thuộc tính Giới tính (và 'Phù hợp để sử dụng' tiếp theo) đáng kể giảm tỷ lệ mất tích **Giới tính** từ **19,5%** ít hơn **0,1%**. Hãy tưởng tượng sự gia tăng hiệu quả trước/sau của các sáng kiến Tiếp thị (hoặc Phân tích xu hướng Phân tích)

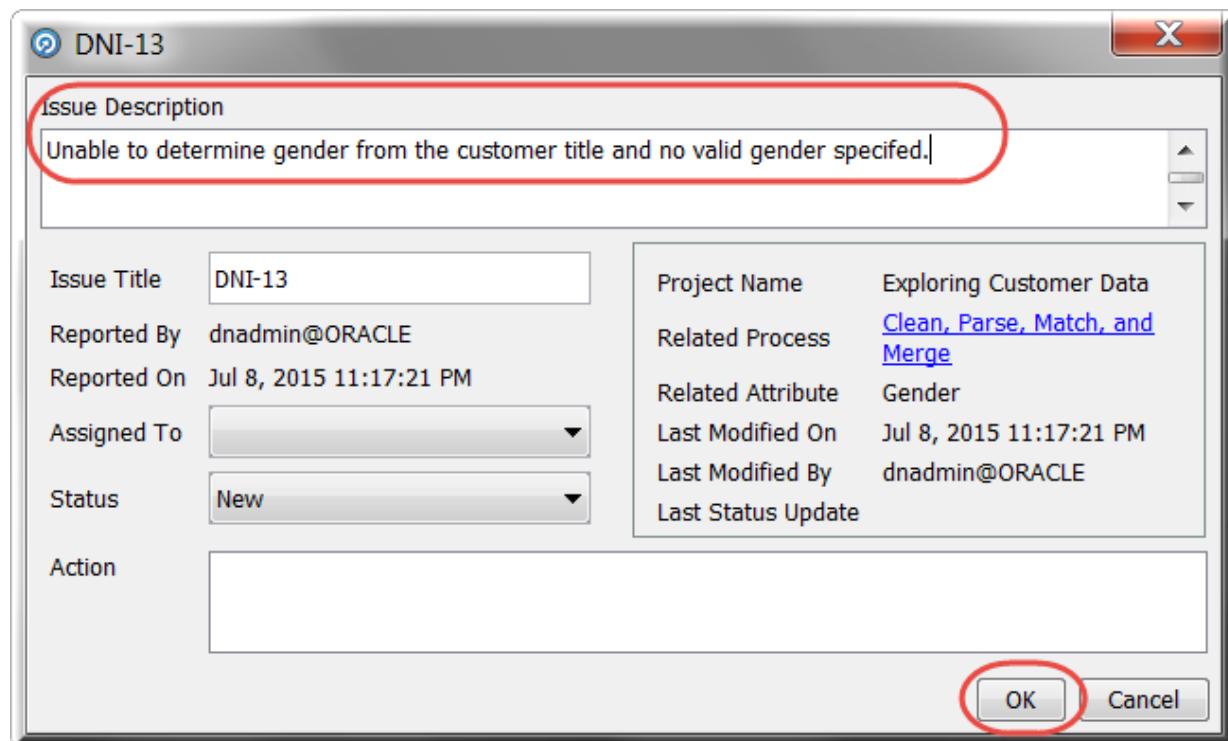
40. Bằng cách quan sát những điều sau đây **Tên**cột, bạn có thể thấy lý do tại sao **Giới tính** không được bắt nguồn

BestGender	TitleGender	NameGender	DerivedGender	ID	Name
				VWR459115	A & S Bowen
				QSN402369	Lessie Sanchez
				ANJ609945	Ryan Arter & Emily May
				AQM457231	Ira Dudley
				ZJS456032	Haydee Humphrey
				BVZ446137	Dr D Nairne-Smith

41. Cuộn thêm về bên phải cho đến khi bạn nhìn thấy **Giới tính**cột.

Lưu ý có những ký tự không hợp lệ. Hãy tạo một vấn đề và quay lại vấn đề đó trong Lab 4. Hãy nhớ rằng, Vấn đề có thể được chỉ định cho người dùng hoặc nhóm. Chúng ta sẽ làm việc nhiều hơn với Vấn đề trong vài phút nữa.

42. Nhấn phím CTRL và nhấp vào **Bạn** giá trị dưới đây **Giới tính**cột. Nhấp chuột phải và nhấp **Tạo vấn đề...**Kiểu **Không thể xác định giới tính từ tiêu đề khách hàng và không có giới tính hợp lệ được chỉ định** cho **Mô tả vấn đề**.

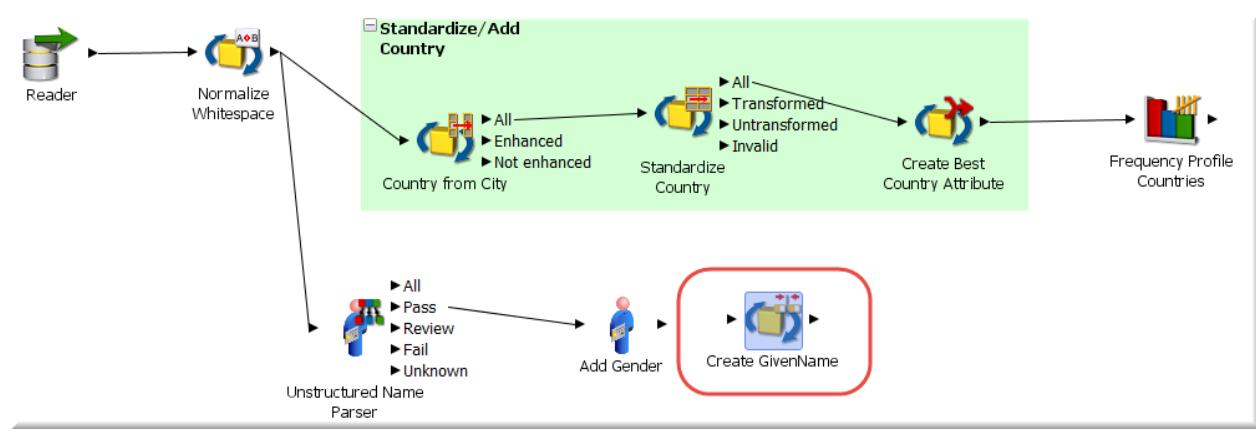


Phòng thí nghiệm 4: So khớp, hợp nhất và loại bỏ dữ liệu trùng lặp để tạo Nguồn dữ liệu duy nhất cho các thực thể dữ liệu chính của bạn

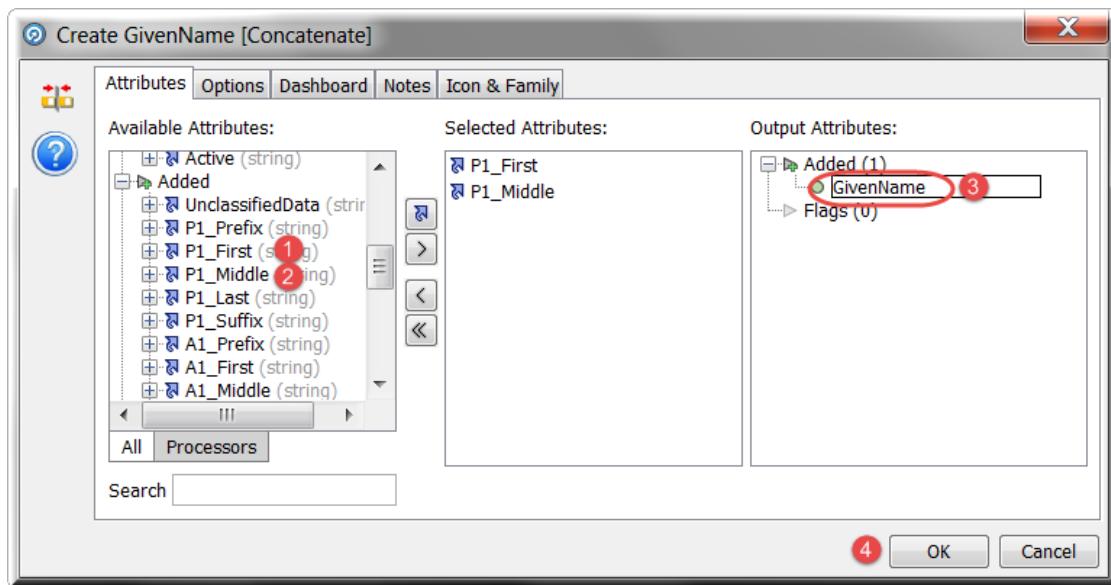
Bây giờ chúng ta đã Làm giàu và Chuẩn hóa các cột Quốc gia và Giới tính (vẫn còn các vấn đề dữ liệu khác – đó là những vấn đề chúng ta tập trung vào trong Phòng thí nghiệm 3), chúng ta sẽ bắt đầu chuẩn bị dữ liệu để đối sánh và sau đó chạy bộ xử lý đối sánh được cấu hình sẵn để đối sánh, hợp nhất và loại bỏ dữ liệu trùng lặp.

1. Quay lại **Bảng công cụ** và tìm kiếm **Nối liền** bộ xử lý bằng cách gõ **Nối liền**

2. Kéo và thả **Nối liền** bộ xử lý trên Project Canvas. Nhấp đúp vào nhãn **Nối liền** để đổi tên bộ xử lý thành **Tạo GivenName**



3. Kết nối tam giác cuối từ **Thêm Giới tính** ĐẾN **Tạo GivenName**. Chọn các thuộc tính **P1_Đầu tiên** Và **P1_Giữa**. Hãy chắc chắn chọn các thuộc tính này theo thứ tự này. Nhấp đúp vào **Concat** trong **Thuộc tính đầu ra** và đổi tên nó thành **Tên đã cho**, nhấp chuột **ĐƯỢC RỒI** để tiếp tục



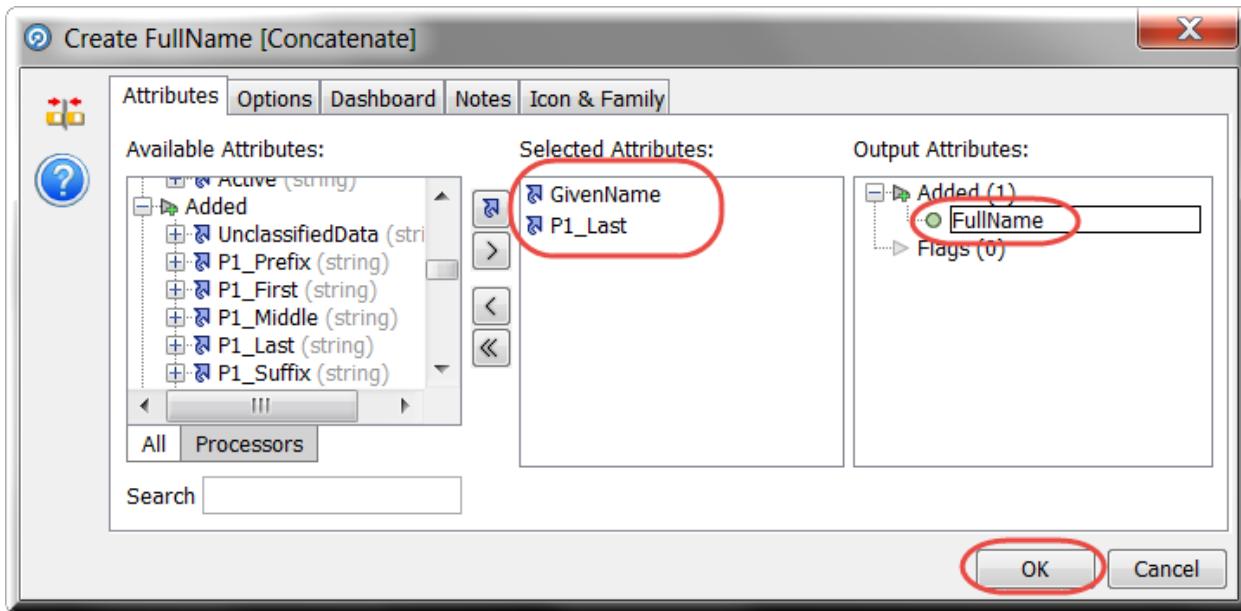
4. Nhấp chuột phải vào **Tạo GivenName** bộ xử lý và nhấp chuột **Sao chép**

5. Nhấp chuột phải vào Project Canvas và nhấp vào **Dán** để thêm một bản sao khác của **Tạo GivenName** bộ xử lý. Nhấp đúp vào tên của bộ xử lý đã sao chép để đổi tên thành **Tạo FullName**. Tiếp theo, kết nối Bộ xử lý **Tạo GivenName** ĐẾN **Tạo Tên Đầu**

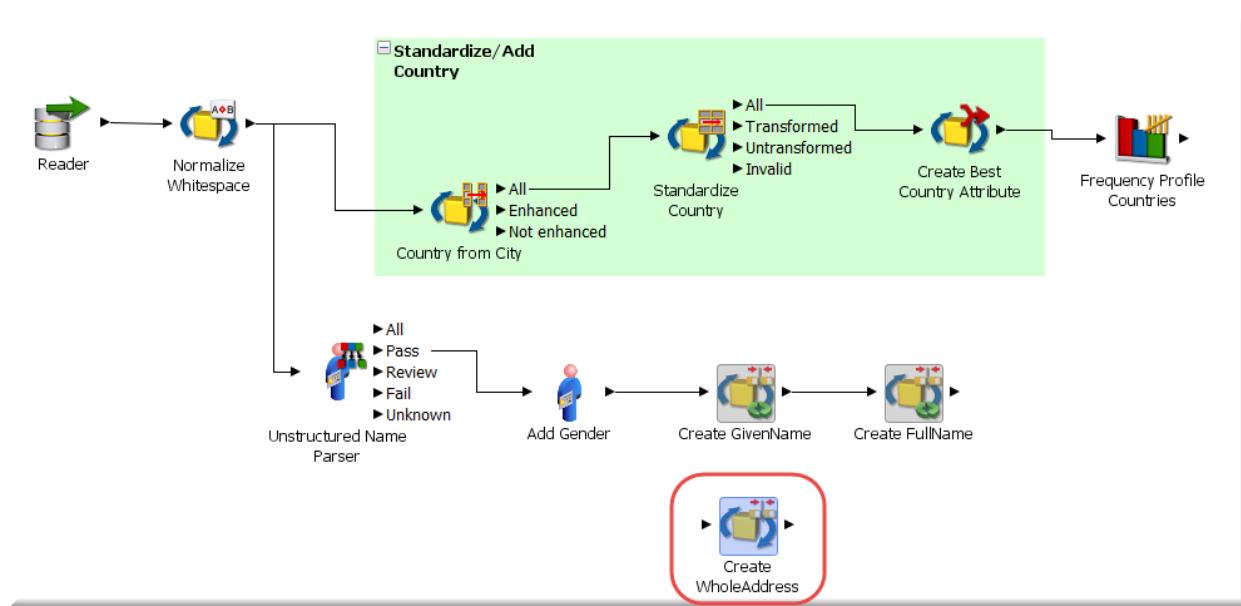


6. Kết nối tam giác cuối của **Tạo GivenName** đến tam giác đầu vào của **Tạo FullName**. Đầu tiên, xóa bỏ hiện tại **Thuộc tính được chọn** bằng cách nhấp đúp vào **Tên đã cho** và **P1_Cuối cùng**

7. Chọn **Tên đã cho** và **P1_Cuối cùng** theo thứ tự này. Đổi tên **Thuộc tính đầu ra** ở phía bên phải của hộp thoại để **Họ và tên đầy đủ** bằng cách nhấp đúp vào **Thuộc tính đầu ra hiện tại**



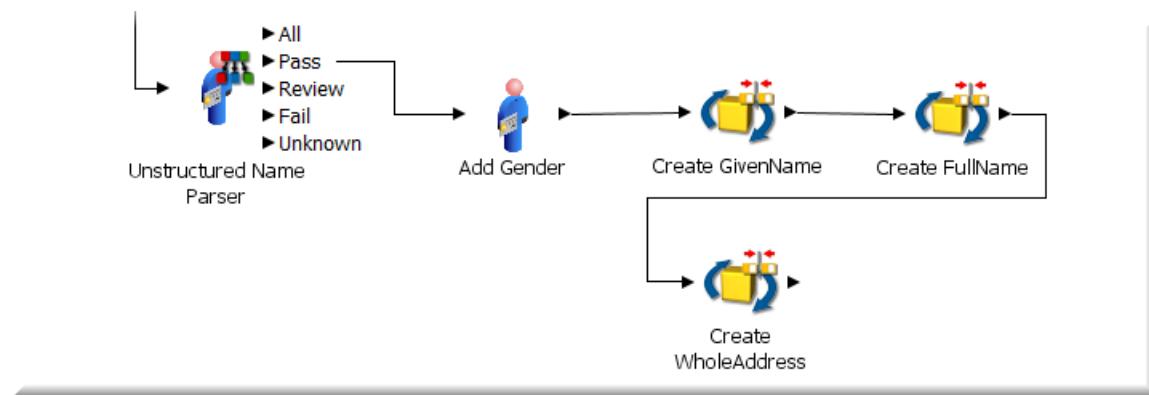
8. Quay lại **Bảng công cụ** và thêm một cái nữa **Nối liền** bộ xử lý cho Project Canvas bên dưới hiện tại **Tạo GivenName** và **Tạo FullName** Bộ xử lý. Nhấp đúp vào tên bộ xử lý để đổi tên thành **Tạo WholeAddress**



Bạn có thể nhấp đúp vào đường ống kết nối các bộ xử lý khác nhau để thay đổi thành kết nối khuỷu tay.

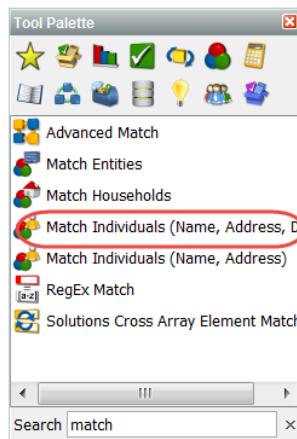
9. Kết nối điểm cuối của **Tạo FullName** vào đầu vào của **Tạo WholeAddress** bộ xử lý. Chọn **Đường phố, Thành phố, Tiểu bang, Và ZIP** các thuộc tính nhu **Thuộc tính được chọn**. Đổi tên **Thuộc tính đầu ra** bằng cách nhấp đúp vào giá trị hiện có để **Toàn bộ địa chỉ**, nhấp chuột **ĐƯỢC RỜI** để tiếp tục

10. Nhấp đúp vào **đường kẻ** kết nối các **Tạo FullName** và **Tạo WholeAddress** bộ xử lý để thay đổi nó thành kết nối khuỷu tay

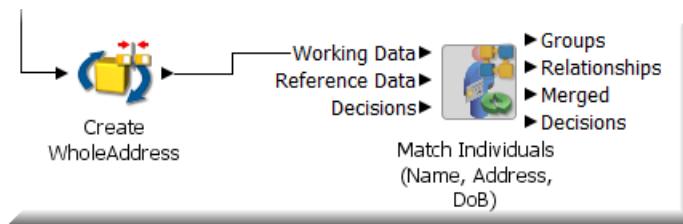


Tiếp theo, chúng ta sẽ tìm và thêm Bộ xử lý phù hợp để Phù hợp với Cá nhân.

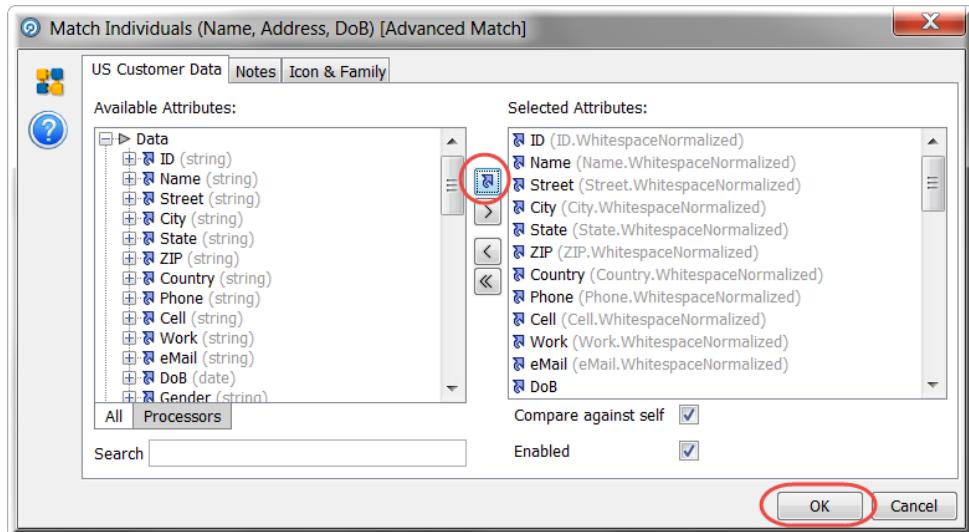
11. Điều hướng đến **Bảng công cụ** và tìm kiếm bằng cách nhập **Cuộc thi đấu**. Bạn có thể cần phải mở rộng **Bảng công cụ** để tìm **Phù hợp với cá nhân (Tên, Địa chỉ, Ngày sinh)**. Nhấp và kéo đường viền bên trái của **Bảng công cụ** để thay đổi kích thước nó



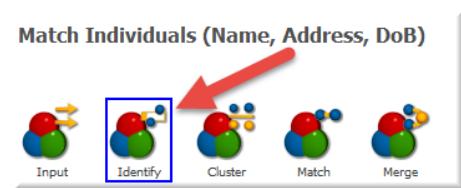
12. Kéo **Phù hợp với cá nhân (Tên, Địa chỉ, Ngày sinh)** vào Project Canvas và kết nối tam giác cuối từ **Tạo WholeAddress** đến **Dữ liệu làm việc** cổng đầu vào của **Phù hợp với cá nhân (Tên, Địa chỉ, Ngày sinh)** Bộ xử lý



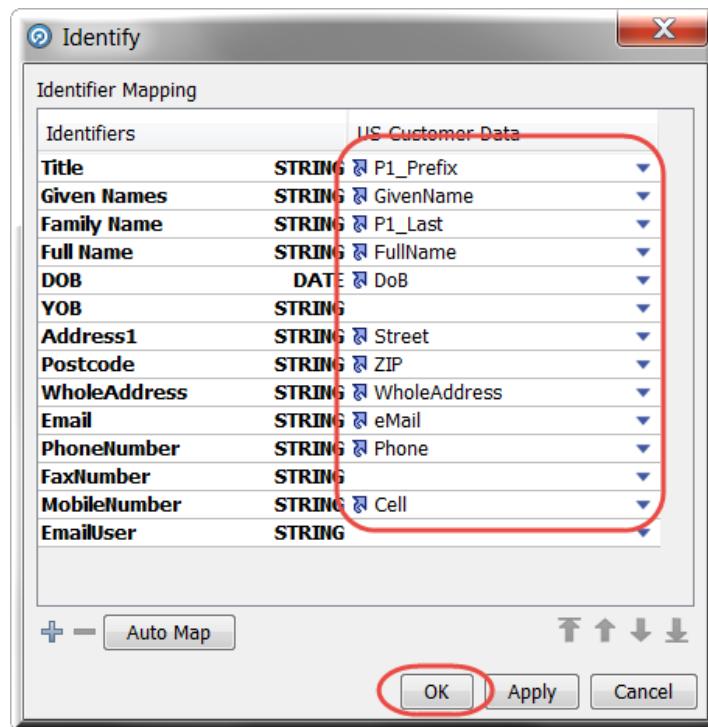
13. Sử dụng **Chọn tất cả** nút để Chọn tất cả các thuộc tính, nhấp vào **ĐƯỢC RỜI** để tiếp tục



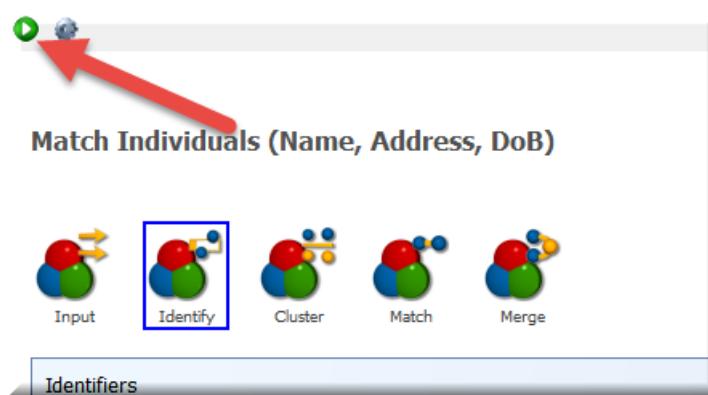
14. Nhấp đúp vào **Nhận dạng** bộ xử lý phụ



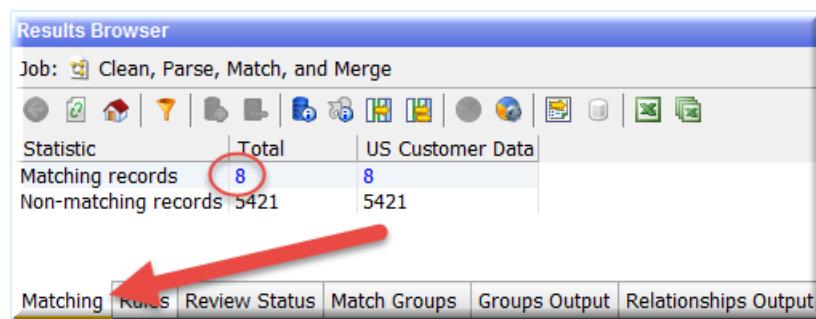
15. Kết nối các thuộc tính đầu vào với các mã định danh phù hợp như được mô tả trong đồ họa ở trang tiếp theo



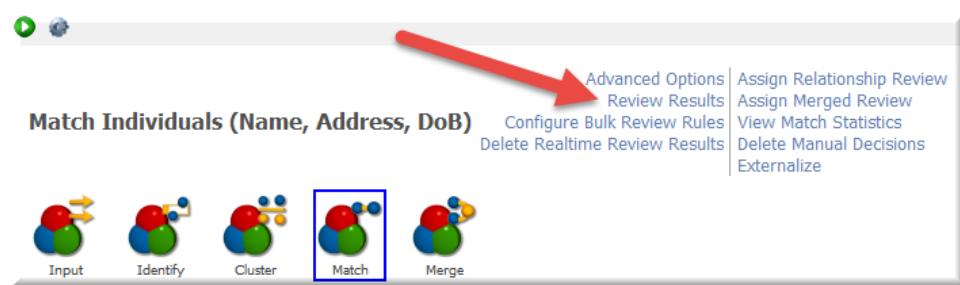
16. Nhấp vào **Chạy** biểu tượng trên thanh công cụ để chạy quy trình



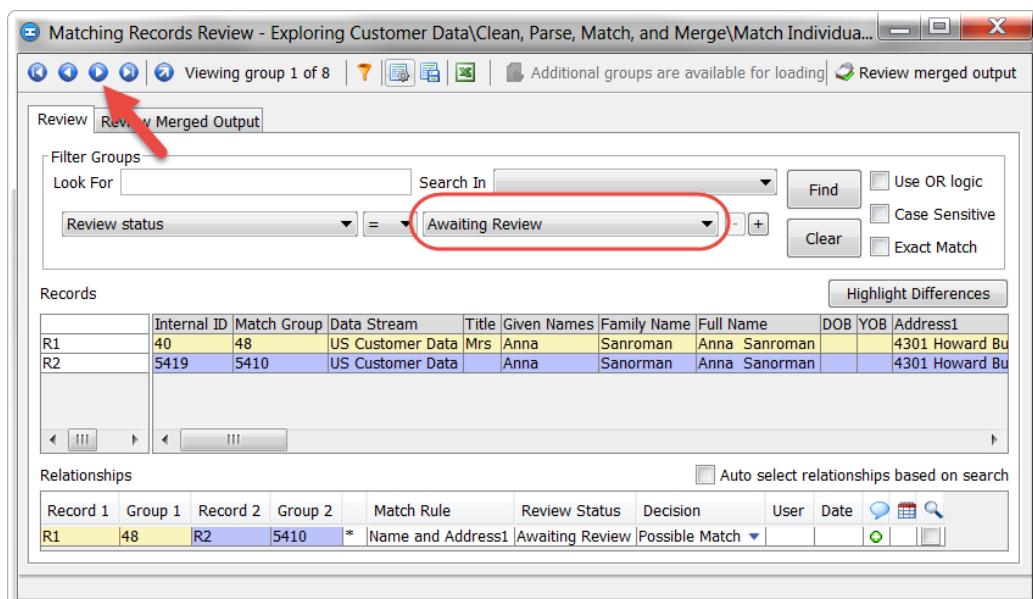
17. Khi quá trình hoàn tất, hãy nhấp một lần vào **Cuộc thi đấu** bộ xử lý phụ để xem kết quả trong **Trình duyệt kết quả**. Chọn **Phù hợp** tab ở cạnh dưới của Trình duyệt kết quả và bạn sẽ thấy Bộ xử lý trận đấu đã tìm thấy **8 hồ sơ** trận đấu tự động



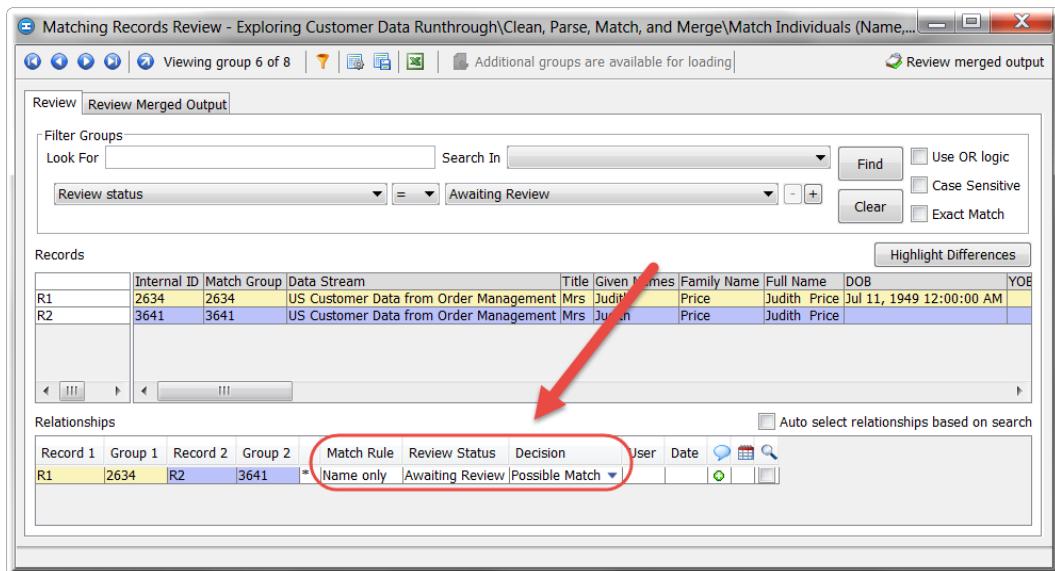
18. Nhấp vào **Đánh giá kết quả** liên kết ở góc trên bên phải của **Dự án Canvas** để xem kết quả khớp nhau chi tiết hơn



19. Sử dụng các nút ở trên cùng bên trái để điều hướng qua các nhóm đang chờ xem xét

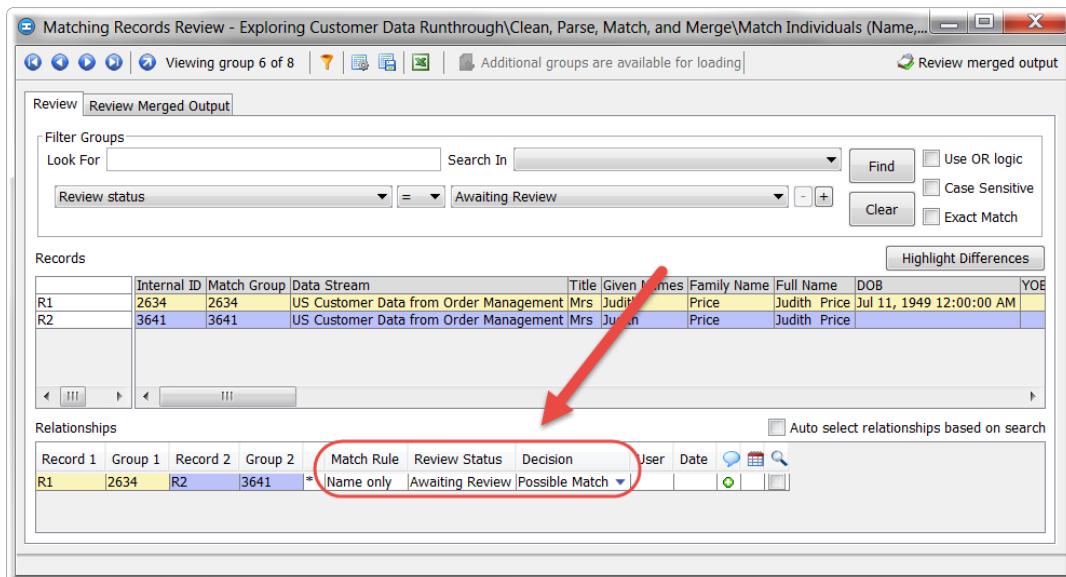


Lưu ý ở phía dưới **Xem lại hồ sơ phù hợp** cửa sổ, có thêm các cột Hiển thị **Quy tắc trận đấu**, **Đánh giá tình trạng**, **Và Phán quyết**.

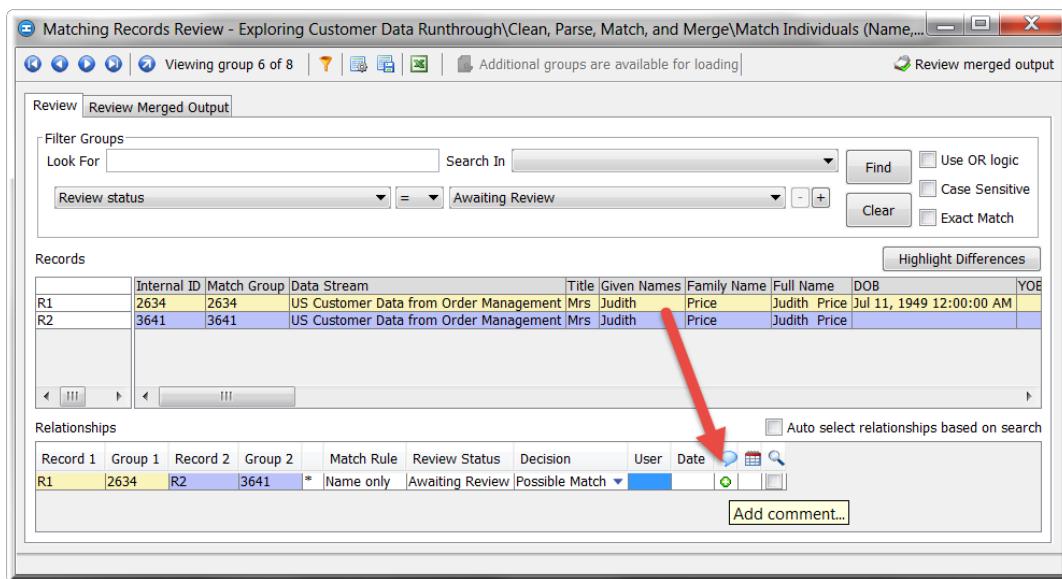


20. Nhập vào Phản quyết thả xuống để xem các tùy chọn có sẵn

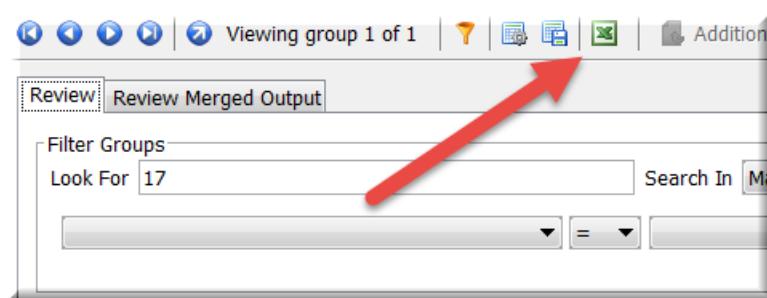
Nếu vấn đề có thể được xác minh bởi chuyên gia về chất lượng dữ liệu, thì một trong những quyết định sau có thể được chọn để khớp với hồ sơ.



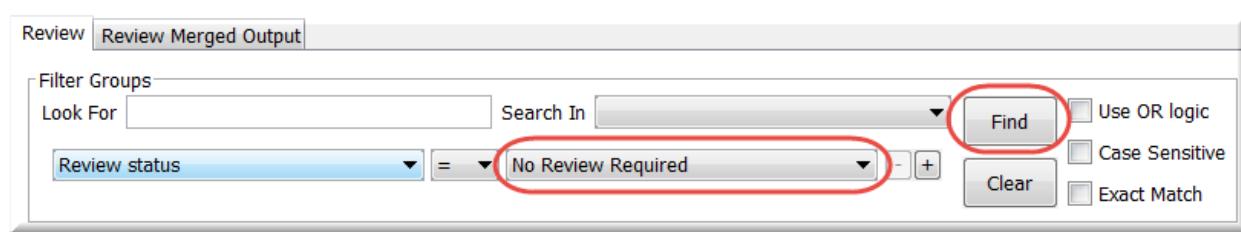
21. Ngoài ra, nếu cần thêm thông tin chi tiết hoặc người dùng chất lượng dữ liệu muốn đưa ra bình luận, **Dấu + Xanh lá cây** ⬤ dưới **Bóng bóng lời thoại** có thể nhấp vào để thêm bình luận



22. Như trong các màn hình Trình duyệt Kết quả trước đó, một **Xuất sang Excel** nút có sẵn ở đây để gửi kết quả Xem xét hồ sơ phù hợp cho người dùng khác để xác minh



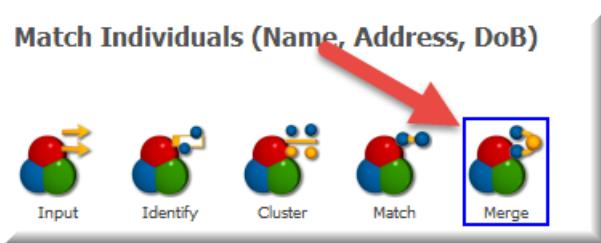
23. Để xem những gì đã được khớp tự động, hãy thay đổi bộ lọc trên màn hình để hiển thị các nhóm có **Trạng thái đánh giá = Không cần đánh giá**



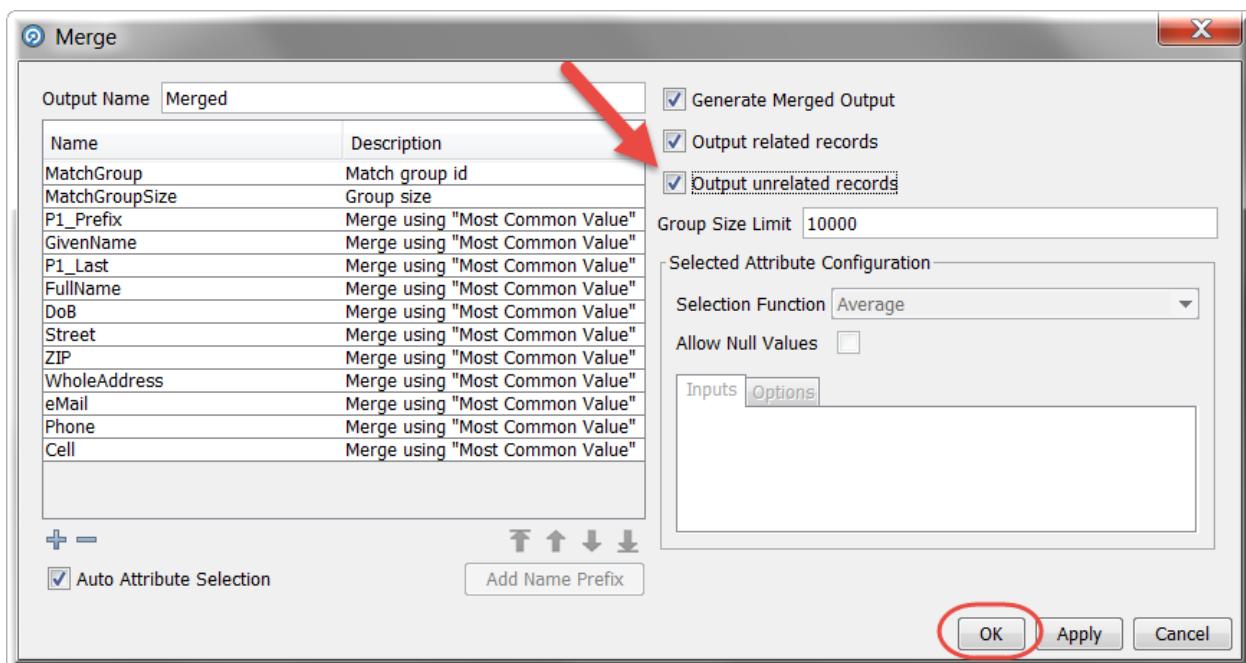
24. Điều hướng qua các nhóm để xem lại kết quả khớp bằng cách sử dụng **Bên trái** và **Phải** mũi tên ở góc trên bên trái của **Xem lại hồ sơ phù hợp** hộp thoại. Khi hoàn tất, hãy đóng **Xem lại hồ sơ phù hợp** cửa sổ

Cho đến nay, chúng ta đã làm sạch, phân tích và khớp dữ liệu từ tập dữ liệu khách hàng này. Tại thời điểm này, chúng ta sẽ sử dụng bộ xử lý phụ Merge để ghi ra các bản ghi đã loại bỏ trùng lặp từ Match Process bằng các quy tắc hợp nhất mặc định.

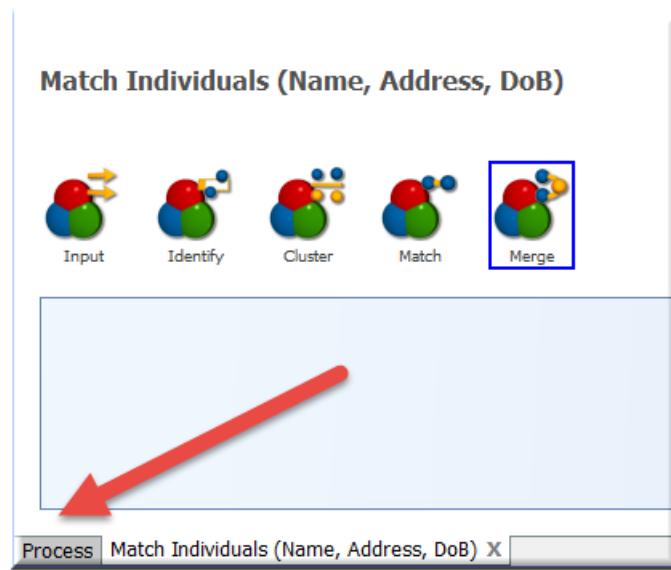
25. Nhấp đúp vào **Hợp nhất** bộ xử lý phụ để bắt đầu và mở **Hợp nhất** hộp thoại



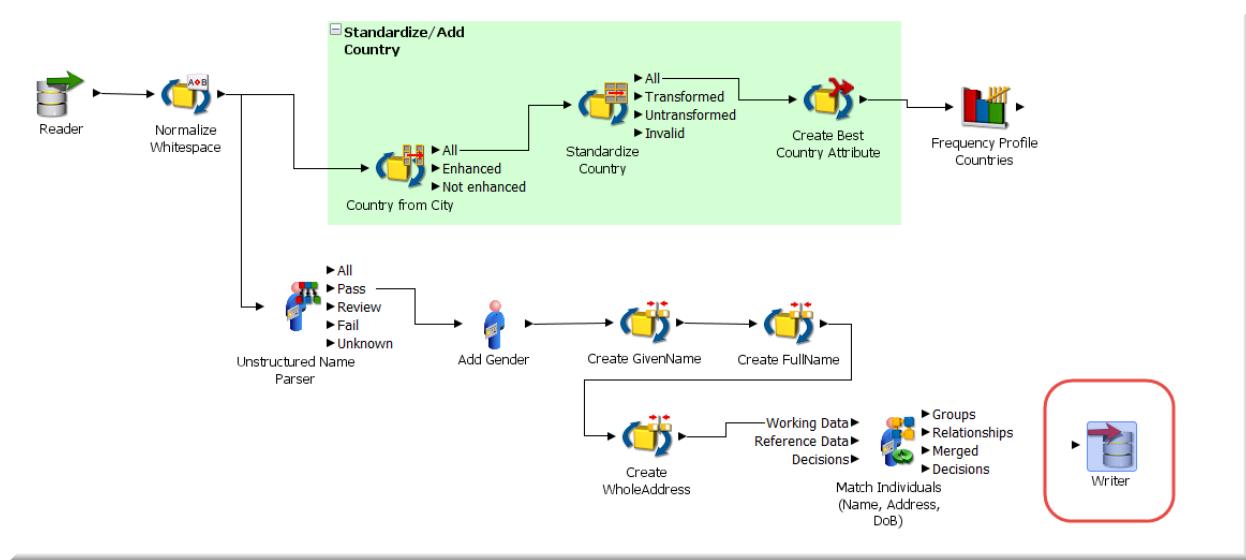
26. Đánh dấu vào ô bên cạnh **Xuất ra các bản ghi không liên quan** trong **Hợp nhất** hộp thoại, nhấp chuột **ĐƯỢC RỒI** để tiếp tục



27. Nhấp vào **Tab quy trình** ở giữa bên trái của canvas để quay lại Project Canvas



28. Trong **Bảng công cụ**, tìm kiếm bằng cách nhập **Nhà văn**. Bộ xử lý này cho phép quy trình EDQ ghi dữ liệu vào các loại kho dữ liệu khác nhau, ví dụ: Dữ liệu được phân giai đoạn. Kéo và thả **Nhà văn** ở bên phải của **Bộ xử lý trận đấu** trên Canvas Dự án

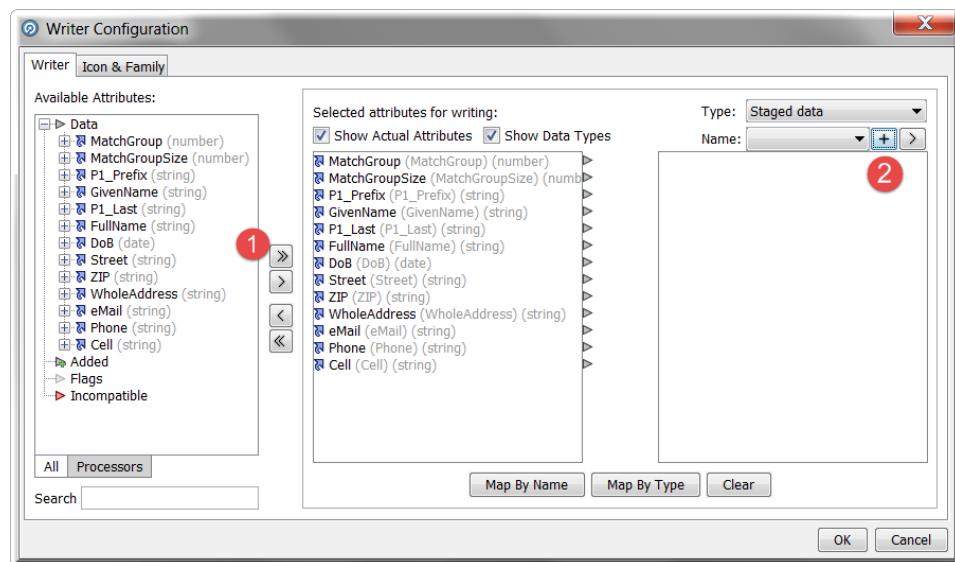


29. Kết nối tam giác cuối từ **Đã hợp nhất** đầu ra từ **Phù hợp với cá nhân** Bộ xử lý cho **Nhà văn**

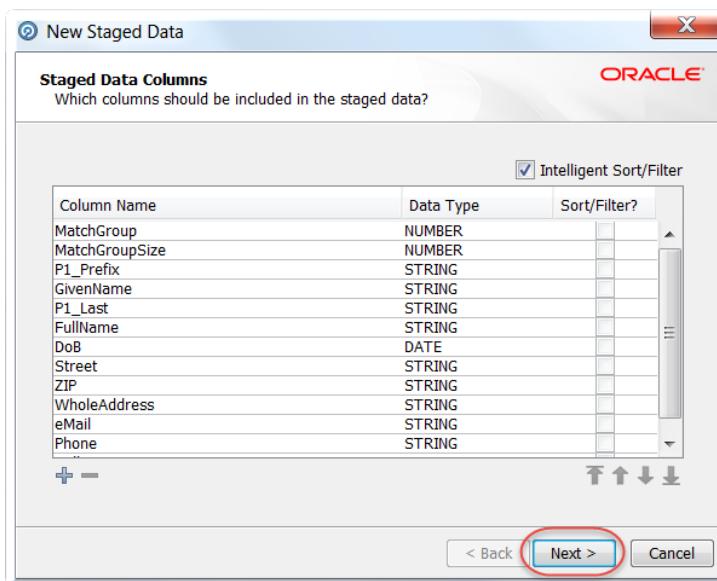


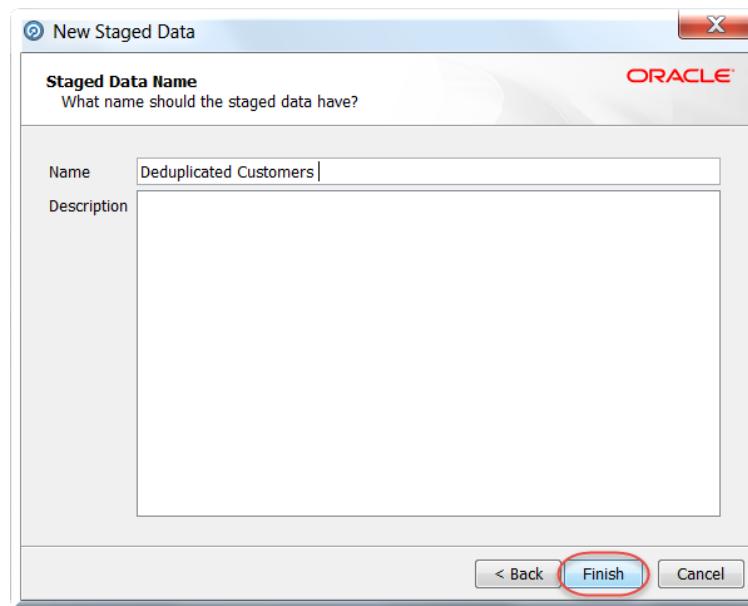
Có những dạng đầu ra khác cho bộ xử lý Match Individuals. Ví dụ, Groups, là các bản ghi đầu vào được sắp xếp thành các nhóm khớp. Relationships, là tất cả các liên kết giữa các bản ghi khớp.

30. Nhấp vào nút để Chọn tất cả **Thuộc tính có sẵn**. Sau đó nhấn nút trên phía bên phải của **Cấu hình Writer** hộp thoại để thêm một cái mới **Dữ liệu được phân loại**

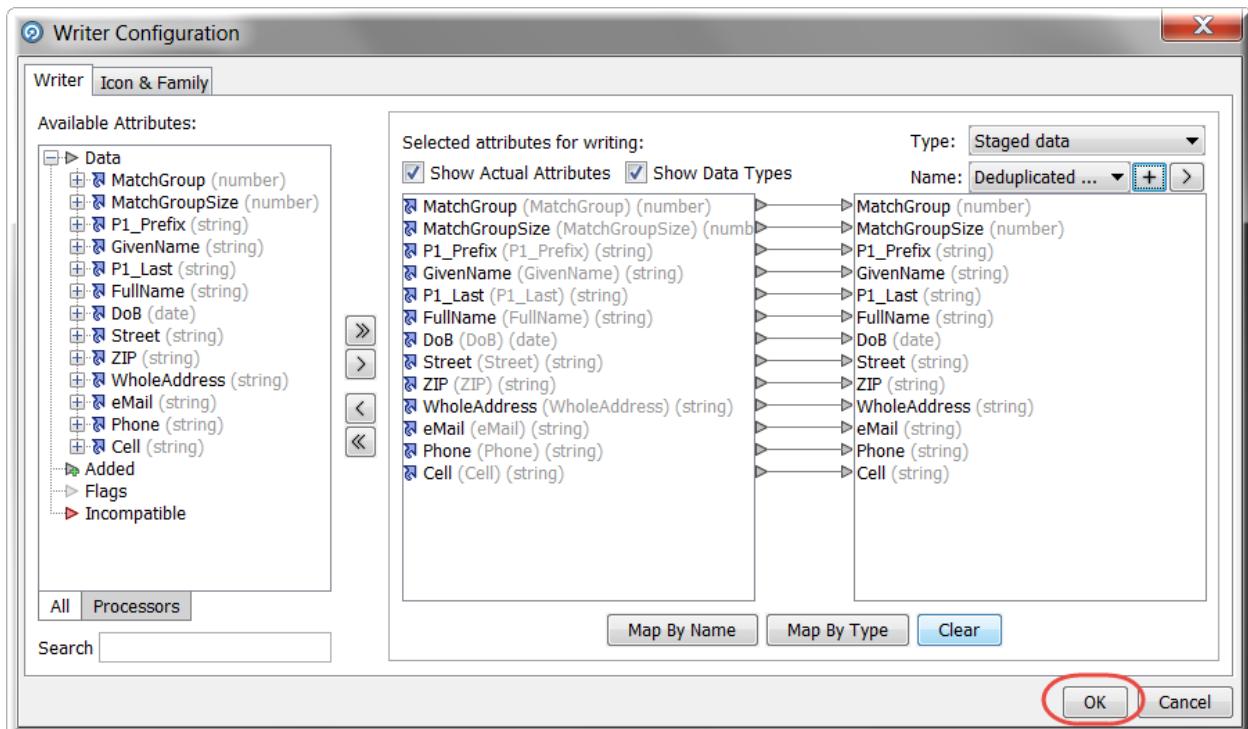


31. Nhấp chuột **Tiếp theo >** để chấp nhận cấu hình mặc định của việc tạo Bộ dữ liệu theo giai đoạn. Đặt tên cho Bộ dữ liệu theo giai đoạn bằng cách nhập **Khách hàng bị loại bỏ trùng lặp**, nhấp chuột **Hoàn thành**





32. Nhấp chuột **ĐƯỢC RỒI** để hoàn tất việc thiết lập **Cấu hình Writer**



33. Nhập vào **Chạy** biểu tượng trên thanh công cụ để chạy quy trình. Khi quy trình này hoàn tất, nhấp chuột phải vào **Nhà văn** và nhấp vào **Hiển thị kết quả trong cửa sổ mới**

MatchGroup	MatchGroupSize	P1_Prefix	GivenName	P1_Last	FullName	DoB	Street
2	2	Mr	Richard	Brown	Richard Brown	Jan 1, 1970 12:00:00 AM	4048 South Lynn Court
1	2	Mr	Brian	Robles	Brian Robles	Nov 1, 1975 12:00:00 AM	10800 Foreman
4	2	Dr	Mudar C	Nairne-Clark	Mudar C Nairne-Clark	Jan 1, 1970 12:00:00 AM	6631 Palmetto Circle S
5	1	Mrs	Eleanor	Peachey	Eleanor Peachey	May 4, 1949 12:00:00 AM	12811 Farmington Road
6	1	Mrs	Christine	Hunt	Christine Hunt	Jan 1, 1970 12:00:00 AM	11197 Leadbeater Road
7	1	Mr	Barry	Ponce	Barry Ponce	May 20, 1962 12:00:00 AM	2660 Auburn Road
8	1	Mr	Walter	Terrell	Walter Terrell	Jan 1, 1970 12:00:00 AM	442 East 18th Avenue
9	1	Mrs	Terri	Smith	Terri Smith	Jul 22, 1968 12:00:00 AM	4100 Park Forest Drive
10	1	Ms	Laurene	Ross	Laurene Ross	Oct 9, 1978 12:00:00 AM	3200 W Euless Boulevard
11	1	Mrs	Virginia	Andrew	Virginia Andrew	Sep 7, 1957 12:00:00 AM	2535 Walnut Hill Lane
12	1	Mr	Lester	Young	Lester Young	Jan 1, 1970 12:00:00 AM	9010 Maier Road
13	1	Mr	Nathan	Hamilton	Nathan Hamilton	Jan 1, 1977 12:00:00 AM	116 Campbell Avenue
14	1	Mr	John	Mcalister	John Mcalister	Jan 1, 1950 12:00:00 AM	5309 Curson Avenue
3	2	Mr	Darren	Hamilton	Darren Hamilton	Jan 1, 1970 12:00:00 AM	480 Equestrian Drive
16	1	Mrs	Phyllis	Ford	Phyllis Ford	Jan 1, 1970 12:00:00 AM	900 Terminal Place

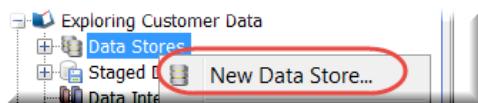


Có nhiều nhóm khớp lệnh khác nhau với MatchGroupSize là 2 - mỗi bản ghi đơn được ghi ra cho 2 bản ghi đầu vào khớp lệnh. Ngoài ra, tổng khối lượng bản ghi được ghi là 5425, thay vì 5438 trong **Người đọc bộ xử lý**.

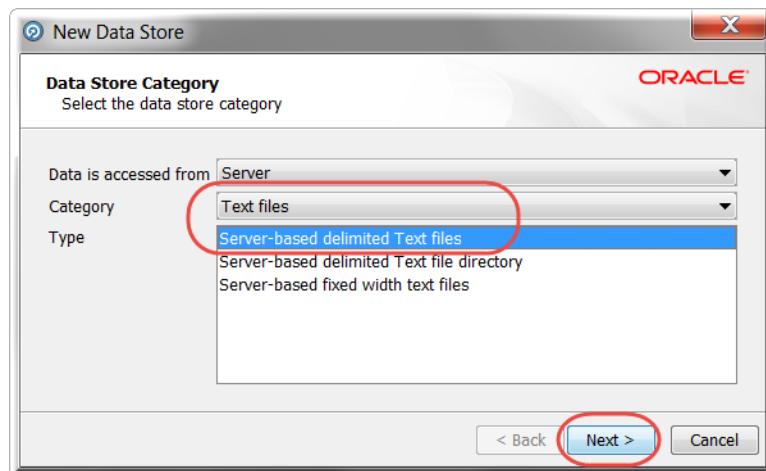
Phòng thí nghiệm 5: Tạo công việc để tự động hóa chất lượng dữ liệu với tích hợp ODI và EDQ

Trước khi chuyển sang Lab 6, chúng ta sẽ tạo một công việc chạy dữ liệu thông qua quy trình chúng ta vừa tạo. Tiếp theo, chúng ta sẽ cấu hình Công việc để chạy trong một gói ODI để tự động hóa dự án Chất lượng dữ liệu.

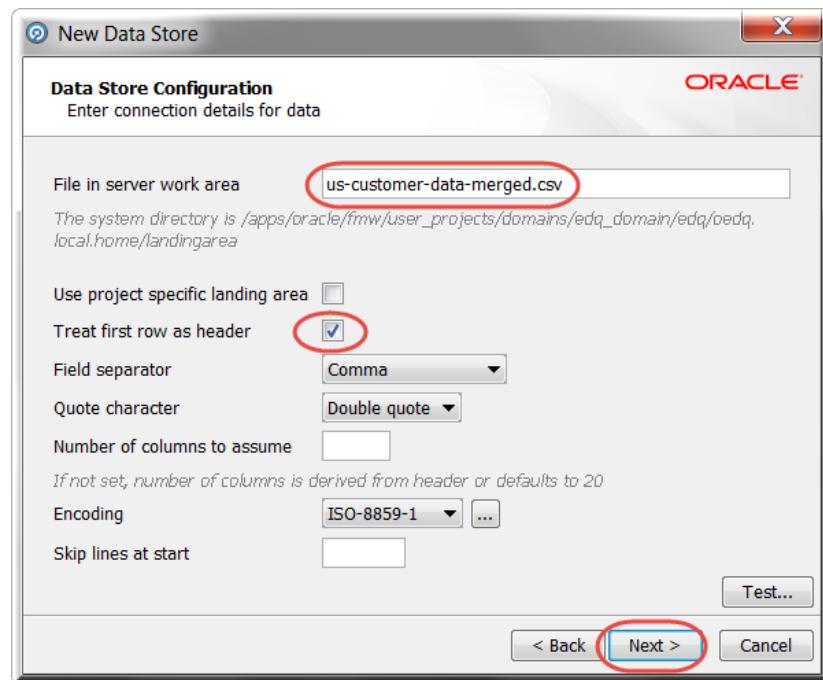
- Trong Trình duyệt Dự án ở phía bên trái màn hình, nhấp chuột phải **Kho dữ liệu** và nhấp vào **Kho dữ liệu mới** dưới của bạn **Khám phá dữ liệu khách hàng** dự án



- Chọn **Máy chủ** và **Tập tin văn bản** trong hai hộp thả xuống, sau đó nhấp vào **Các tập tin văn bản được phân cách dựa trên máy chủ** và nhấp vào **Tiếp theo >** để tiếp tục

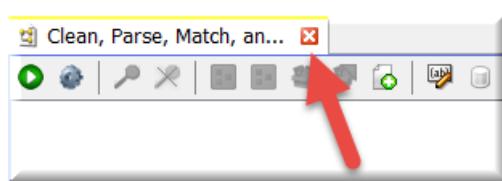


- Trong vùng làm việc Tệp trong máy chủ, nhập tên cho tệp xuất bằng cách nhập **us-customerdata-merged.csv**. Ngoài ra, hãy kiểm tra tùy chọn để **Xử lý hàng đầu tiên như tiêu đề**

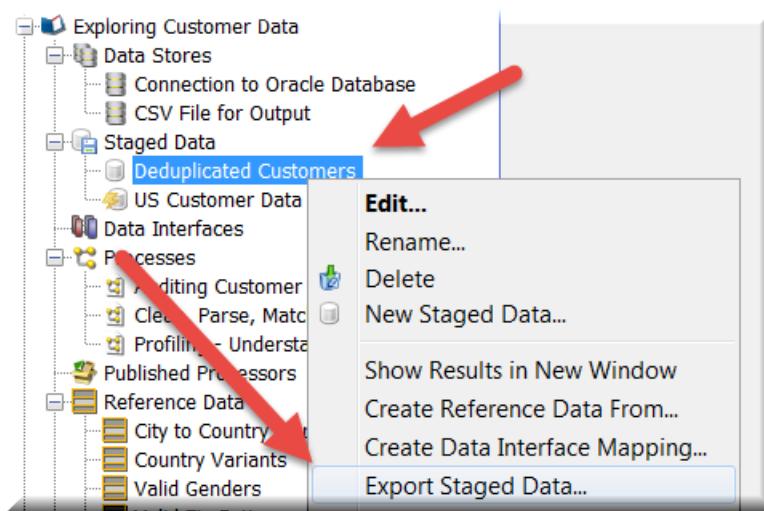


4. Nhấp vào **Tiếp theo** > và đặt tên cho Kho dữ liệu này – loại **Tệp CSV để xuất**, nhấp chuột **Hoàn thành**

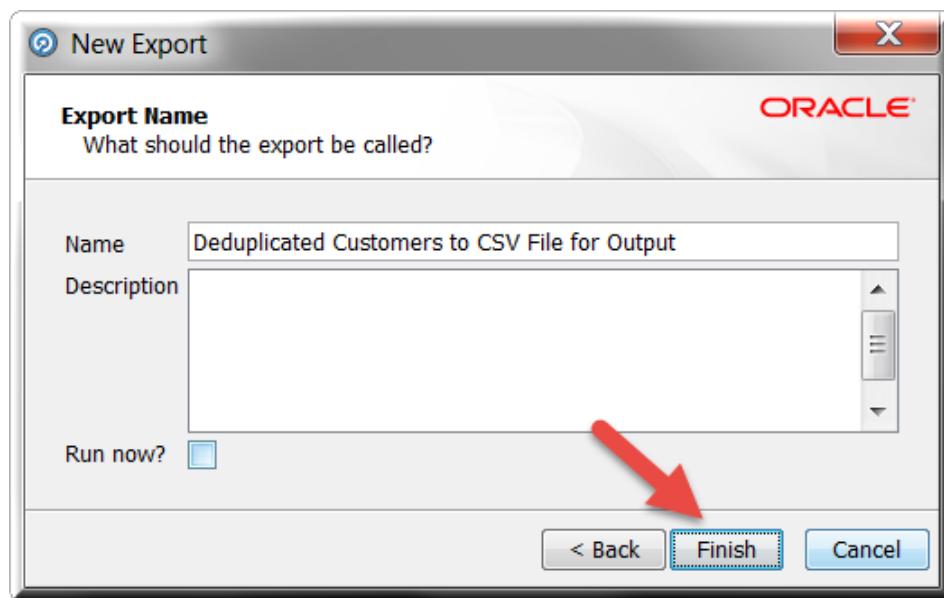
5. Đóng cửa bạn **Các tab quy trình** ở phía trên cùng của **Dự án Canvas** nếu bạn chưa làm. Nếu bạn được nhắc lưu một quy trình nhất định, hãy nhấp vào **Đúng**



6. Điều hướng đến Trình duyệt Dự án ở phía bên trái của màn hình và nhấp chuột phải vào **Khách hàng bị loại bỏ trùng lặp** dưới **Dữ liệu được phân loại** danh mục. Mở rộng danh mục Dữ liệu được dàn dựng nếu bạn không thấy nó. Chọn **Xuất dữ liệu theo giai đoạn...**. Nhấp chuột **Tiếp theo** > trên cửa sổ đầu tiên của hộp thoại vì Dữ liệu được phân giai đoạn đã được chọn



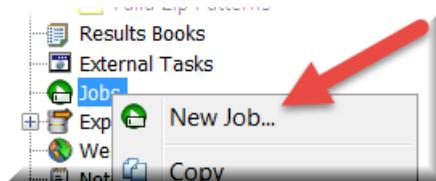
7. Chọn **Tệp CSV để xuất** kho dữ liệu làm mục tiêu cho việc xuất dữ liệu. Nhấp vào **Kế tiếp**
 > **Và Hoàn thành** để giữ tên xuất mặc định và hoàn tất trình hướng dẫn để
 cấu hình Cấu hình xuất mới



Tạo việc làm

Tiếp theo chúng ta sẽ tạo một công việc có thể được gọi từ bên ngoài, ví dụ, từ Oracle Data Integrator.

8. Điều hướng đến Trình duyệt Dự án và chọn **Việc làm** danh mục. Nhấp chuột phải vào nó và chọn **Công việc mới...**

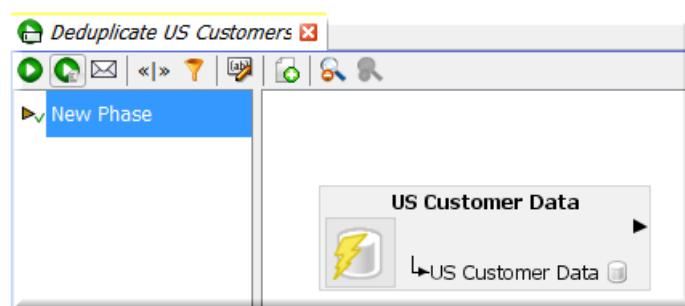


9. Đặt tên cho công việc bằng cách nhập **Loại bỏ trùng lặp khách hàng Hoa Kỳ**, nhấp chuột **Hoàn thành**

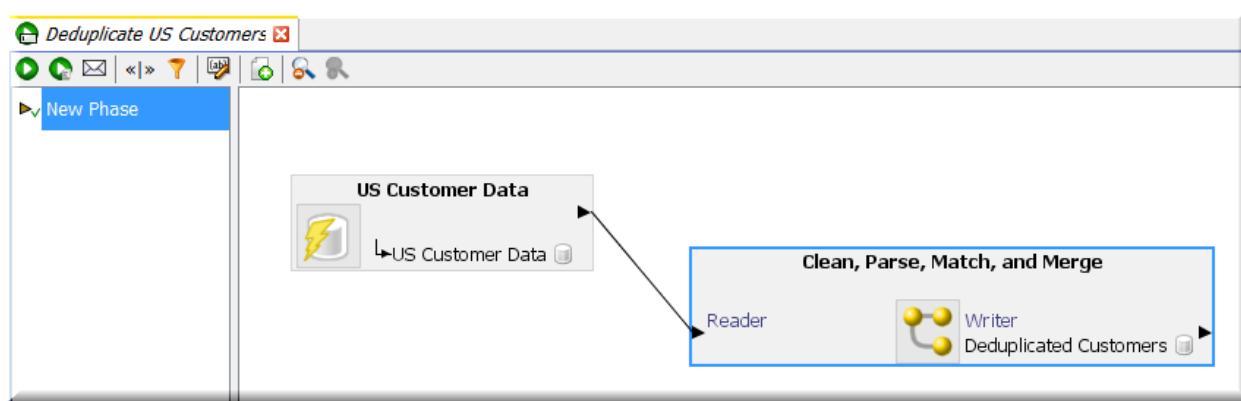
Tiếp theo, Job Canvas được hiển thị với Bảng công cụ hơi khác một chút. **Bảng công cụ** chứa tất cả các tác vụ cấu hình có thể chạy trong dự án bao gồm ảnh chụp nhanh, quy trình, xuất, v.v. Lưu ý **các biểu tượng ở đầu Bảng công cụ** khác nhau – hãy nhấp vào chúng để khám phá.



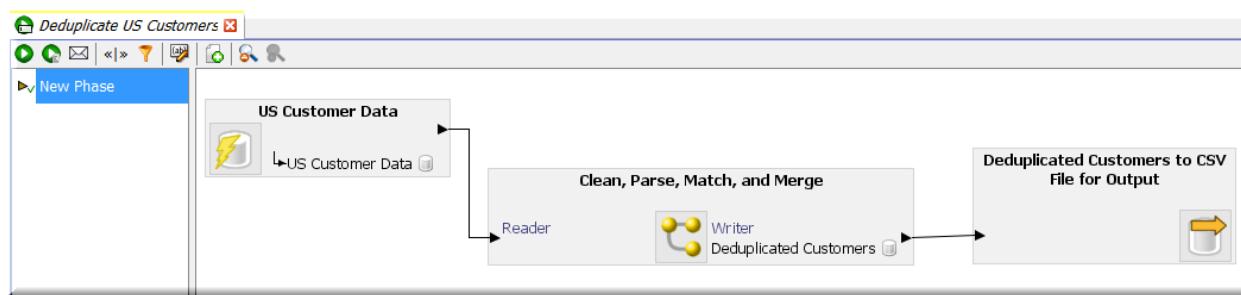
10. Nhấp vào biểu tượng trong **Bảng công cụ** để hiển thị **Ảnh chụp nhanh** và kéo **Kết nối đến Oracle Database.US_Customer_DATA** chụp nhanh vào bức tranh



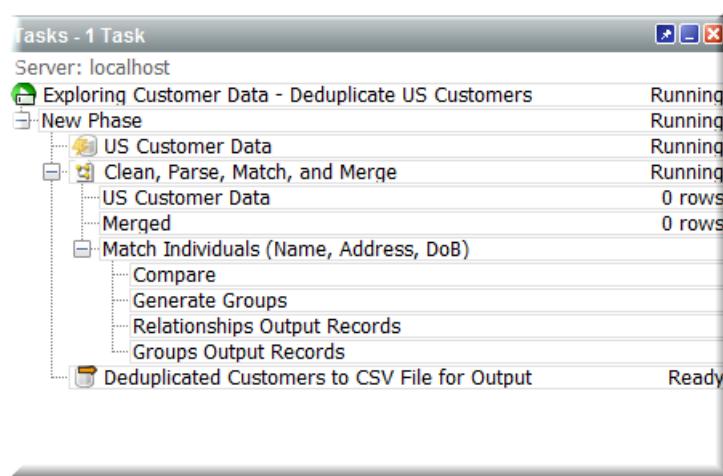
11. Tiếp theo, nhấp vào biểu tượng trong **Bảng công cụ** để hiển thị **Các quy trình** và kéo **Làm sạch, Phân tích, So khớp và Hợp nhất** xù lý lên vải



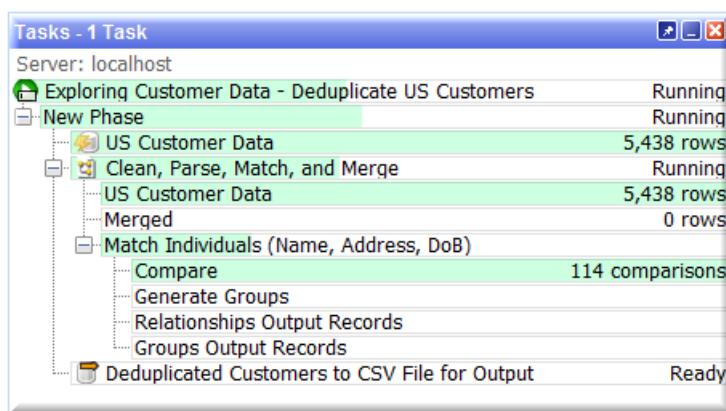
12. Cuối cùng, nhấp vào ở phía bên phải của **Bảng công cụ** để hiển thị **Xuất khẩu**. và thả **Khách hàng trùng lặp vào tệp CSV** để xuất ra bản vẽ



13. Nhấp vào **Chạy**biểu tượng trên thanh công cụ để chạy công việc. Lưu ý **Nhiệm vụ**Cửa sổ ở góc dưới bên trái của Giám đốc



Có ba tác vụ – Ảnh chụp nhanh, Xử lý và Xuất) đều chạy cùng nhau để làm mới dữ liệu đầu vào, chạy dữ liệu thông qua quy trình bạn đã tạo và ghi tệp vào tệp CSV mới trong vùng đích của máy chủ.



Khi công việc chạy xong, bạn có thể nhấp vào **lối tắt đến khu vực hạ cánh EDQ** trên màn hình nền của Máy ảo và xem **us-customer-data-merged.csv** tập tin mà công việc đã tạo ra.

14. Nhấp vào  biểu tượng trên thanh tác vụ của bạn và chuyển sang màn hình nền của EDQ 12.2.1 Máy ảo

15. Nhấp đúp vào  phím tắt và tìm **us-customer-data-merged.csv**

```

MatchGroup,MatchGroupSize,P1_Prefix,GivenName,P1_Last,FullName,DoB,Street,ZIP,WholeAddress,eMail,Phone
2,2,Mr,Lester ,Young,Lester Young,,9010 Maier Road,20723,9010 Maier Road LAUREL MD 20723,,,(534) 405 7444,(344) 719 3606
4,2,Mr,Walter ,Terrell,Walter Terrell,,442 East 144th Avenue,80020,442 East 144th Avenue BROOMFIELD CO 80020,,,(330) 738 2431,(589) 048 1176
5,2,Miss,Phylis ,Ford,Phylis Ford,1958-10-08 00:00:00,900 Terminal Place,23220,900 Terminal Place RICHMOND VA 23220,,,(163) 441 6358,(255) 386 7434
6,2,Mr,Brian ,Robles,Brian Robles,1975-01-11 00:00:00,10800 Foreman,49330,10800 Foreman LOWELL MI 49331,Brian.K.Robles@xtmail.com,(740) 937 3714,
7,2,Mrs,Tasha ,Trower,Tasha Trower,,2300 Ford Street,80401,2300 Ford Street GOLDEN CO 80401,,,(151) 315 5132,
8,2,Mr,Richard ,Brown,Richard Brown,,4048 South Lynn Court,64055,4048 South Lynn Court INDEPENDENCE MO 64055,Richard.S.Brown@scene46.com,(462) 347 6367,(741) 915 4823
9,2,Mr,Perry ,Seibert,Perry Seibert,,202 C Street,92101,202 C Street SAN DIEGO CA 92101,,,(644) 6935745,
10,2,Mr,Darren ,Hamilton,Darren Hamilton,,480 Equestiran Drive,75032,480 Equestiran Drive ROCKWALL TX 75032,,,(121) 303 9378,(185) 455 2754
12,2,Mr,Ignacio ,Harris,Ignacio Harris,,4800 Harvey Street,49444,4800 Harvey Street MUSKEGON MI 49444,,,(134) 388 2292,(128) 335 7287
1,2,Mrs,Chris ,Hunt,Chris Hunt,1985-06-05 00:00:00,11197 Leadbetter Rd,23005,11197 Leadbetter Rd ASHLAND VA 23005,Christine.G.Hunt@snomail.com,(739) 636 0289,
3,2,Mr,J ,McAlister,J McAlister,1950-01-01 00:00:00,5309 Curson Avenue,93309,5309 Curson Avenue BAKERSFIELD CA 93309,,(843) 731 2627,(417) 745 6296
11,2,Dr,Mudar C,Nairne-Clark,Mudar C Nairne-Clark,1979-01-07 00:00:00,6631 Palmetto Circle S,33433,6631 Palmetto Circle S BOCA RATON FL 33433,Carol.J.Neale@thu.com,(764) 016 6737,(533)

```

Công việc này hiện có thể được sử dụng với trình lập lịch tích hợp của Oracle Enterprise Data Quality, bất kỳ lịch trình bên ngoài nào hoặc sử dụng tác nhân tích hợp sẵn với Oracle Data Integrator để tạo Kho dữ liệu phù hợp để sử dụng và Phân tích chính xác.

Tích hợp ODI và EDQ – Theo dõi (*Không có bài tập thực hành*)

Trong ví dụ ngắn này, bạn sẽ thấy cách tự động hóa các dự án Data Quality và cách tích hợp một công việc với ODI. Vì ODI không được cài đặt trong môi trường EDQ workshop, các mô tả và ảnh chụp màn hình sau đây nhằm mục đích cho bạn thấy một ví dụ về cách tích hợp giữa hai sản phẩm có thể được thực hiện.

Vì có một số thành phần và khái niệm khác liên quan đến việc tích hợp ODI và EDQ, nên điều quan trọng là phải đưa ra một số kiến thức cơ bản. Có thể tạo một số ánh xạ trong ODI để thực hiện một số loại luồng dữ liệu và chuyển đổi khác nhau. Các ánh xạ được tạo trong ODI có thể được sử dụng trong các gói, cho phép người dùng ODI tạo luồng và thứ tự mà các ánh xạ đó sẽ thực thi. Ngoài việc bao gồm các ánh xạ trong các gói, ODI cũng

bao gồm Công cụ mở có thể được sử dụng trong các gói để giới thiệu các hành động bổ sung như gửi email, truy xuất tệp bằng FTP hoặc phát tiếng bíp khi gói thành công hay thất bại.

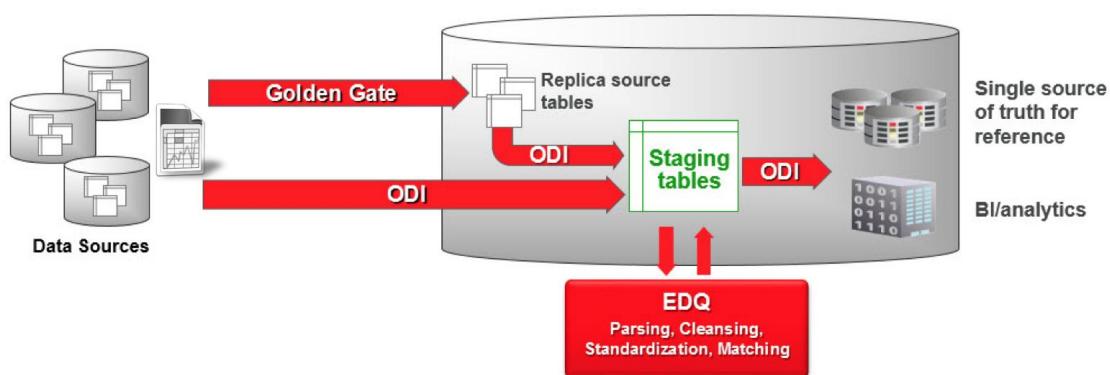
Một Công cụ Mở bổ sung đi kèm với ODI được gọi là Công cụ Mở EnterpriseDataQuality. Công cụ này có thể được sử dụng để gọi các tác vụ EDQ từ một gói ODI bằng giao diện Java Management Extension (JMX). Sau khi kéo công cụ này vào một gói ODI, một vài tham số sẽ cần được chỉ định.

Có thể có nhiều loại trường hợp sử dụng khác nhau để tích hợp ODI với EDQ.

Typical Use Case for EDQ and Data Warehouses

- New/Upgrade to Data Warehouse
- New/Upgrade to Data Mart
- Source data is suspect

- Clean, consistent, usable data
- Accurate analytics
- Single source of truth



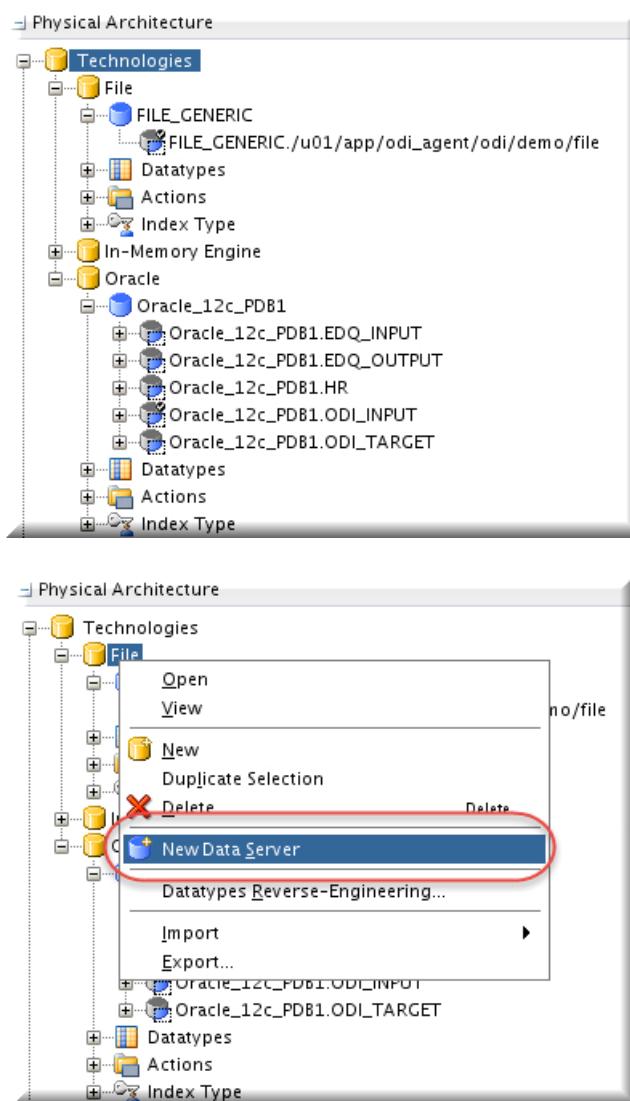
Trong ví dụ sau, chúng ta sẽ tạo một ánh xạ ODI để tải tệp .CSV vào bảng cơ sở dữ liệu, đây là nguồn Đầu vào ODI được sử dụng để tải bảng Đầu vào EDQ. EDQ sẽ đọc Đầu vào EDQ để xử lý bằng Công việc EDQ, giống như công việc đã tạo trước đó. Công việc này sẽ lập hồ sơ, kiểm tra, phân tích cú pháp, làm sạch, khớp và hợp nhất cũng như loại bỏ dữ liệu trùng lặp như chúng ta đã thực hiện trong các phòng thí nghiệm trước.

Sau khi quá trình EDQ hoàn tất, EDQ sẽ ghi đầu ra của công việc trở lại bảng cơ sở dữ liệu, mà chúng ta sẽ gọi là Đầu ra EDQ. Cuối cùng, ODI Mapping sẽ tải Đầu ra EDQ vào bảng Mục tiêu ODI. Ví dụ, Mục tiêu ODI có thể là một kho dữ liệu hoặc kho dữ liệu được sử dụng để phân tích.

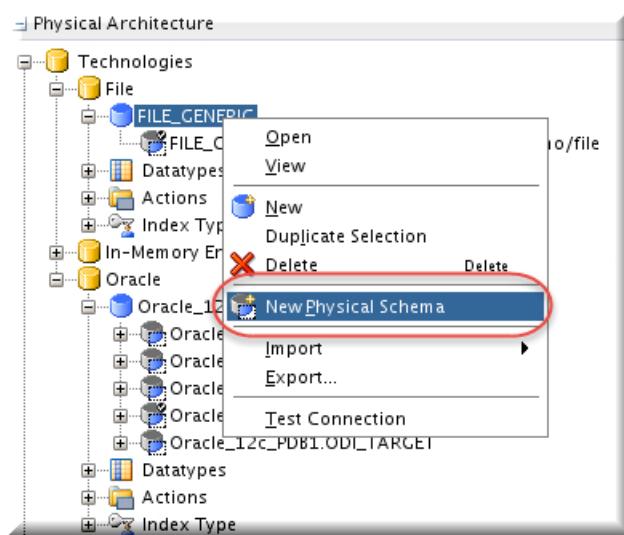
Thiết lập cấu trúc ODI mẫu

Trong ví dụ này, chúng tôi sẽ thiết lập một số kho dữ liệu trong ODI Topology bằng cách sử dụng File và Oracle Physical Technologies.

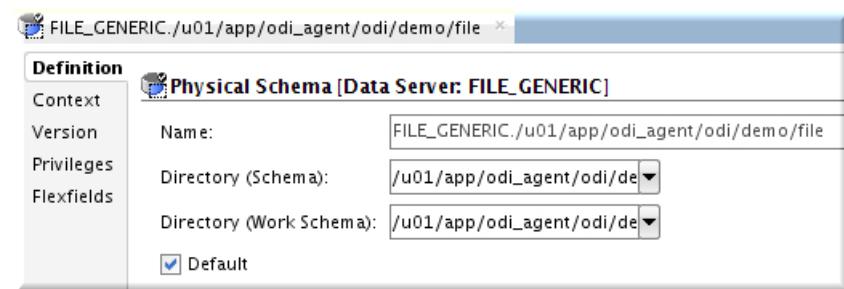
- Để thực hiện di chuyển dữ liệu từ tệp đến bảng phân loại cơ sở dữ liệu đến mục tiêu cơ sở dữ liệu, các máy chủ dữ liệu và kho dữ liệu sau đây cần được thiết lập. Nhấp chuột phải vào các công nghệ, ví dụ **Tài liệu** và **Tiền trù** được liệt kê bên dưới cho thấy tùy chọn để tạo



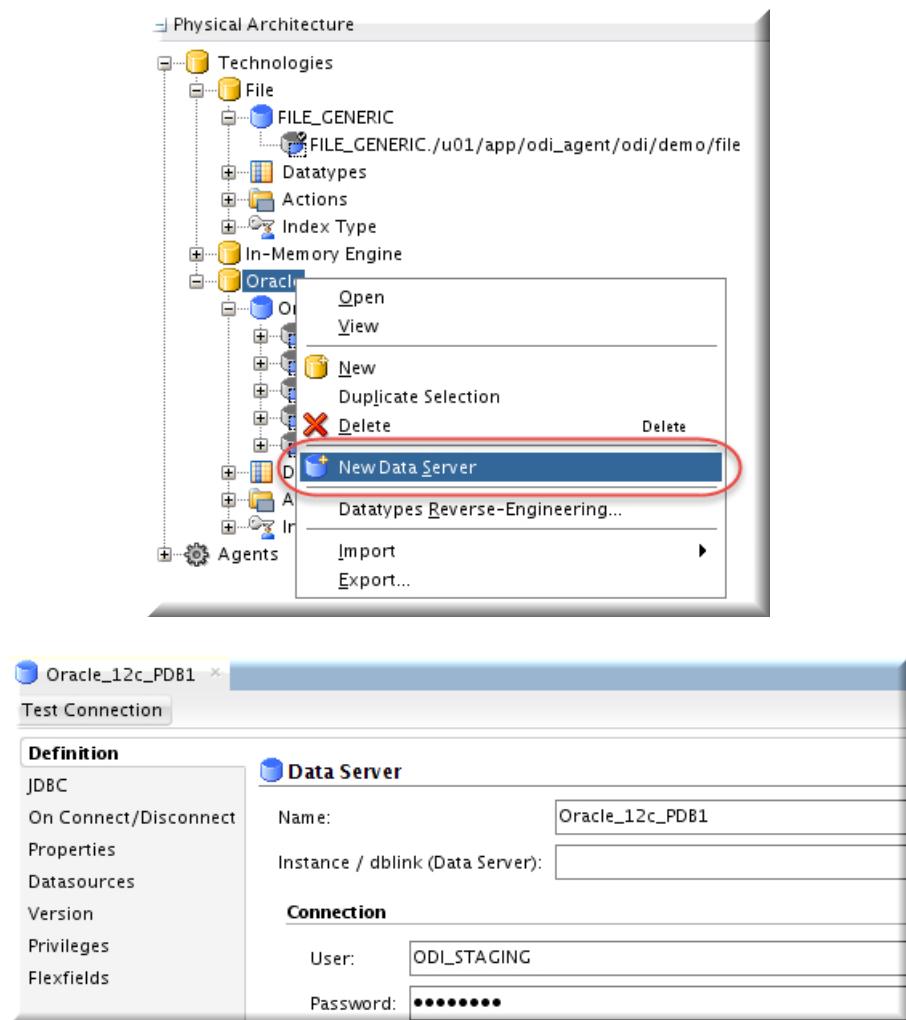
- Trong trường hợp **Tài liệu**, các giá trị mặc định cho **TỆP GENERIC** máy chủ dữ liệu được sử dụng. Nhấp chuột phải vào **TỆP GENERIC** máy chủ dữ liệu tiết lộ các tùy chọn để tạo ra một **Sơ đồ vật lý mới**



3. Sơ đồ vật lý được tạo ra như mô tả trong ảnh chụp màn hình sau. Trong trường hợp **Thư mục (Sơ đồ)** và **Thư mục (Sơ đồ công việc)**, vị trí của tệp .CSV chứa dữ liệu khách hàng đã được cung cấp: `/u01/ứng dụng/odi_agent/odi/demo/tệp`. Tệp . CSV này được lưu trữ trên cùng bộ nhớ với máy khách đang chạy ODI Studio

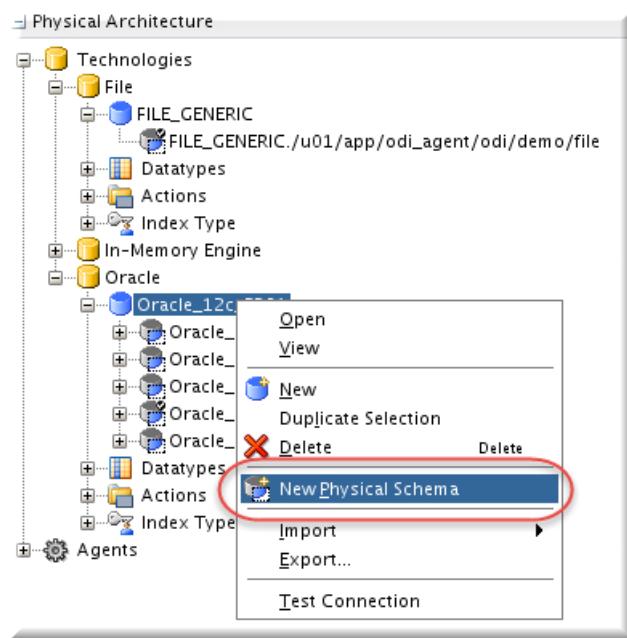


4. Tiếp theo, các kết nối vật lý tới Cơ sở dữ liệu Oracle sẽ cần được tạo để tạo điều kiện thuận lợi cho việc di chuyển dữ liệu từ tệp .CSV sang Bảng đầu vào ODI và các bảng phân đoạn tiếp theo. Một **Tiên tri** kết nối máy chủ dữ liệu được tạo bằng cách nhấp chuột phải vào **Tiên tri** công nghệ

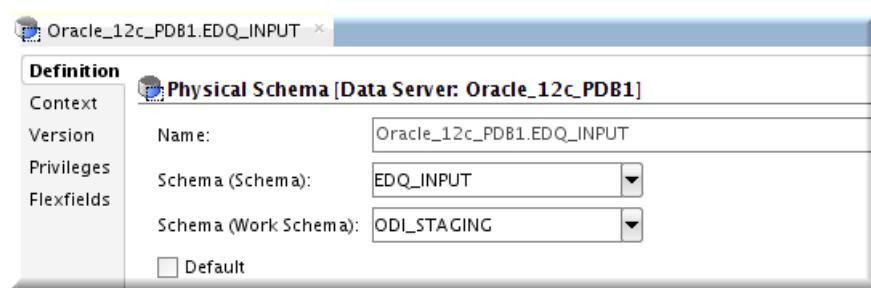


Chuỗi kết nối tới cơ sở dữ liệu Oracle này đã được cung cấp trên **JDBC** tab. Sau đó, bạn có thể nhấp vào **Kiểm tra kết nối** ở góc trên bên trái của màn hình cấu hình kho dữ liệu. Sau khi kết nối thành công, bạn có thể bắt đầu tạo cấu hình cho kho dữ liệu (sơ đồ) để kết nối.

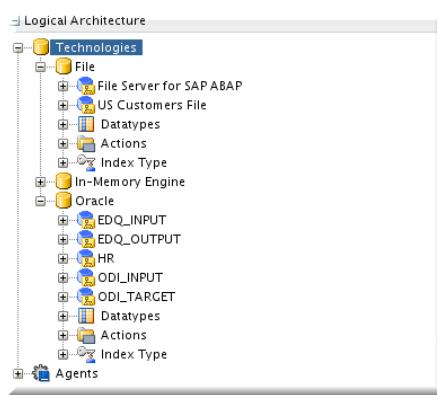
5. Ví dụ này sẽ yêu cầu 4 cấu hình lưu trữ dữ liệu khác nhau: **Đầu vào ODI**, **Đầu vào EDQ**, **Đầu ra EDQ**, Và **Mục tiêu ODI**. Các bảng đầu vào và đầu ra sẽ đóng vai trò là bảng trung gian để thực hiện công việc kiểm tra chất lượng dữ liệu nhằm chuẩn bị tải dữ liệu lên bảng mục tiêu.
6. Nhấp chuột phải vào **Lưu trữ dữ liệu** bên dưới **Tiến trình** trong tab Topology – Physical Architecture cho phép thiết lập một **Sơ đồ vật lý mới**



7. Để giảm sự lặp lại, cả bốn lược đồ vật lý sẽ được tạo bằng các bước giống hệt nhau. Sự khác biệt duy nhất sẽ là lược đồ được chọn trong **Sơ đồ (schema)** thả xuống. Mỗi lựa chọn sẽ tham chiếu đến các lược đồ được tạo trước đó có tên là Đầu vào ODI, Đầu vào EDQ, Đầu ra EDQ và Mục tiêu ODI



8. Cũng giống như thiết lập Topology ODI thông thường, một lược đồ logic phải được tạo cho mỗi kho dữ liệu được tạo ở trên. Nhấp chuột phải vào các công nghệ **Tài liệu** và **Tiên trisẽ** để tạo **Sơ đồ logic mới**. Các lược đồ logic sau đây đã được tạo ra

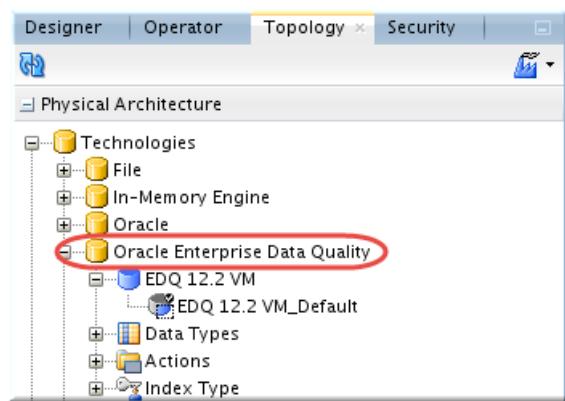


9. Đối với mỗi lược đồ logic được tạo ra, **Sơ đồ vật lý** đã được chọn vào **Sự định nghĩa tab**

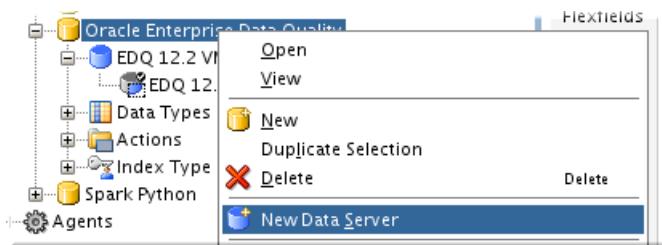
The screenshots show the 'Logical Schema' definition tab for four objects:

- US Customers File**: Name: US Customers File, Physical Schemas: FILE_GENERIC/u01/app/odi_agent/odi/demo/file
- EDQ_INPUT**: Name: EDQ_INPUT, Physical Schemas: Oracle_12c_PDB1.EDQ_INPUT
- EDQ_OUTPUT**: Name: EDQ_OUTPUT, Physical Schemas: Oracle_12c_PDB1.EDQ_OUTPUT
- ODI_INPUT**: Name: ODI_INPUT, Physical Schemas: Oracle_12c_PDB1.ODI_INPUT

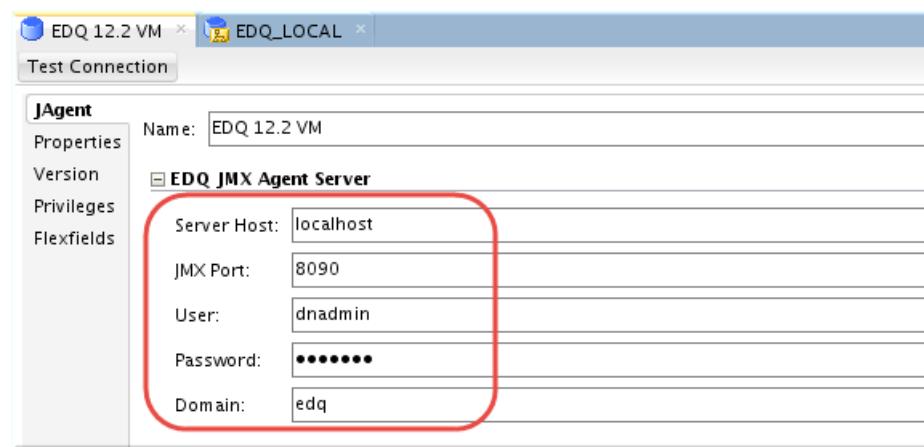
10. Trong Oracle Data Integrator 12.2.1, Enterprise Data Quality là một tùy chọn công nghệ trong Topology Navigator. Một máy chủ dữ liệu được tạo ra để lưu trữ thông tin chi tiết về máy chủ EDQ để sử dụng trong Gói ODI.



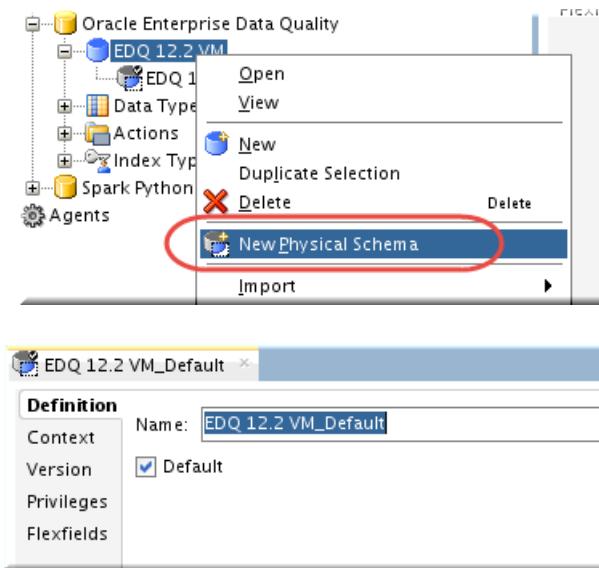
11. Nhấp chuột phải vào **Chất lượng dữ liệu doanh nghiệp Oracle** và lựa chọn **Máy chủ dữ liệu mới** hiển thị hộp thoại để thiết lập kết nối máy chủ EDQ:



12. Một **Tên** cho máy chủ dữ liệu EDQ và **Máy chủ đại lý EDQ JMX** chi tiết được điền như sau. 8090 là cổng JMX mặc định được cấu hình trên VM EDQ này.



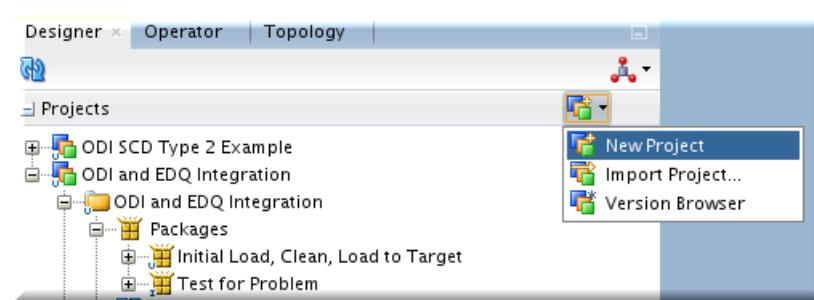
13. Sau đó, một lược đồ vật lý được tạo ra và lưu với thông tin mặc định



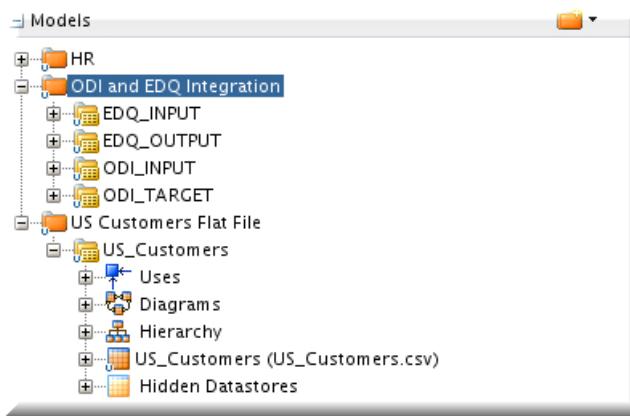
14. Cũng giống như với máy chủ dữ liệu tệp và cơ sở dữ liệu, một Sơ đồ logic đã được tạo cho Máy chủ dữ liệu vật lý EDQ. Chi tiết này sẽ được cung cấp sau khi sử dụng Công cụ mở EDQ trong Gói ODI



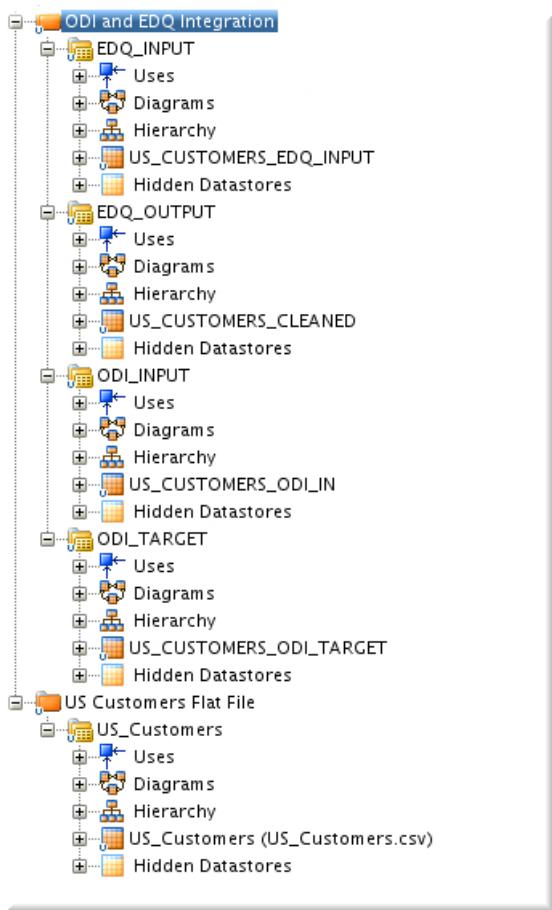
15. Một **Dự án** được tạo ra trong **Nhà thiết kế** tab để giữ các mô-đun kiến thức, ánh xạ và gói cho tích hợp chất lượng dữ liệu mẫu này



16. Thư mục mô hình và các mô hình sau đây đã được tạo trong **Nhà thiết kế** tab và **Mô hình phàn** của ODI



17. Kỹ thuật đảo ngược đã được hoàn thành để tạo ra các mô hình cần thiết để tạo ảnh xạ trong ODI



18. Nhấp chuột phải vào **Khách hàng Hoa Kỳ (US_Customers.csv)** mô hình cho phép bạn xem dữ liệu thô mà chúng ta sẽ làm sạch trong các bước sau. Nhấp vào **Xem dữ liệu** để xem dữ liệu thô

Data: US_Customers

ID	NAME	STREET	CITY	STATE	ZIP	COUNTRY	PHONE	CELL	WORK	EMAIL	
1	AGW727896	Mrs Barbara Saleh	350 East International Speedway	DELAND	FL 32724	(261) 633 4776				Barbara.W.Saleh@shockmail.com	03/11/1
2	ANA572184	Mr Richard Toles	568 Colony Street	Meredith	CT 06450	United States	203-723-7850			Richard.B.Toles@shockmail.com	10/01/1
3	ANL505391	Mrs Esther Tinsley	2626 South Raritan Circle	ENGLEWOOD	CO 80110	USA	(557) 005 3390				01/29/1
4	ARP718985	Mrs Marilyn Lane	600 E Hurst Boulevard	HURST	TX 76053	USA	(272) 993 9897	(483) 779 6735		Marilyn.K.Lane@snomail.com	03/29/1
5	ASW719257	Mrs Margaret Johnson	2104 Union Avenue	SHEBOYGAN	WI 53081	USA	(536) 245 6814			Margaret.C.Johnson@shockmail.com	01/29/1
6	AWJ610448	Mrs Amalia Royall	535 Tanglewood Road	Jackson	MS 39201	U.S.	813-776-7983	287-764-1384	315-677-8883		05/01/1
7	AXY664112	Mr Bernard Trammel	745 Joseph Street	New Berlin	WI 53151		432-646-8770	076-465-3241	236-646-7640	Bernard.O.Trammel@xmail.com	03/28/1
8	BAA645993	Mr Stephen Drake	4978 Mulberry Lane	West Palm Beach	FL 33410		315-623-1029	226-238-1539	519-326-2838		03/11/1
9	BHH630097	Mr John Morris	8020 River Stone Drive	FREDERICKSBURG	VA 22407	USA	(366) 675 0588	(120) 328 2624	(363) 238 5992	John.A.Morris@thu.com	04/30/1
10	BJH586819	Mr Denny Maxwell	518 Pollard Road	ABILENE	TX 79602	USA	(846) 876 4265	(726) 433 0685		Denny.C.Maxwell@shockmail.com	01/01/1
11	BRD597474	Mr Rodolfo Carroll	670 Powers Drive	EL DORADO HILLS	CA 95762	USA	(771) 161 1300	(310) 777 9642		Rodolfo.J.Carroll@snomail.com	09/12/1
12	BRN565104	Mrs Ilene Cribb	11420 Lackland Road	SAINT LOUIS	MO 63146	USA	(446) 647 2594				02/25/1
13	BWM690629	Mr Daniel Castillo	7901 N Tiffany Springs Plaza	KANSAS CITY	MO 64153	U.S.	(804) 174 3541	(402) 656 4638	(319) 476 1166	Daniel.L.Castillo@snomail.com	03/20/1
14	CAC515445	Mr Scotty Butler	990 Lassen Lane	EL DORADO HILLS	CA 95762	USA	(241) 474 6687	(672) 212 3748	(799) 017 2841	Scotty.P.Butler@thu.com	03/05/1
15	CEP464787	Mr Aldo Dozier	1 Church Lane	LOCUST GROVE	VA 22500	USA	134			Aldo.B.Dozier@snomail.com	08/07/1
16	CER446528	Mrs Bebe Myers	8270 US 42	FLORENCE	KY 41042	USA	(243) 938 2678	(505) 675 7710	(181) 912 2727		06/02/1
17	CNN717151	Mrs Karen Lindemann	1160 Caprice Drive Unit C	CASTLE ROCK	CO 80109		(730) 407 7770			Karen.D.Lindemann@xmail.com	11/01/1
18	CSJ638503	Mrs Jean Beckett	777 West Burning Tree Drive	KANSAS CITY	MO 64145	USA	(792) 258 2680	(121) 854 6292	(277) 517 0741	Jean.R.Beckett@xmail.com	09/17/1
19	DDL575105	Mrs Sonya Macias	100 S Mitchell Road	MANSFIELD	TX 76063	USA	no calls			Sonya.D.Macias@xmail.com	05/30/1
20	DDU542204	Mr Raymond Brown	518 Kenilworth Lane	BALLWIN	MO 63011		(363) 633 0336	(347) 322 8615		Raymond.B.Brown@shockmail.com	06/01/1
21	DEV681103	Mrs Melissa Ramey	4802 Valley View Boulevard	ROANOKE	VA 24012	U.S.A.	(867) 665 8377	(474) 514 4829		Melissa.C.Ramey@thu.com	05/11/1
22	DJC685684	Mr Ignacio Harris	4800 Harvey Street	MUSKEGON	MI 49444	United States	(134) 388 2292	(128) 335 7287			09/18/1
23	DNU545353	Mrs Janet Maldonado	6136 West Marconi Avenue	GLENDALE	AZ 85306	United States	(795) 523 0994	(849) 087 3170		Janet.R.Maldonado@snomail.com	05/08/1
24	DNZ481841	Mr Christopher Santora	8101 Ralston Road	ARVADA	CO 80002		(693) 464 0985	(669) 034 7280	(122) 112 7637		11/30/1
25	DZV606283	Mrs Rosalia Willett	11200 SW 49 Place	FORT LAUDERDALE	FL 33330	US	(178) 214 2641				01/01/1
26	ECR636634	Mr Robert Shurel	3242 South Platte River Drive	ENGLEWOOD	CO 80110	USA	(660) 697 1745			Robert.I.Shurel@thu.com	01/13/1
27	EEQ638216	Mrs Doris Parmenter	3301 Rider Trail South	EARTH CITY	MO 63045		(812) 481 1802				07/05/1
28	EFR542975	Mrs Kathryn Thomas	5656 FM 773	VAN	TX 75790	USA	(736) 465 3815	(857) 563 7401			10/12/1
29	EY705465	Mr Thomas Gandara	21412 North 11th Avenue	PHOENIX	AZ 85027	United States	(695) 391 9206				09/28/1
30	EQY445677	Mrs Shirley Gonzalez	1401 McWilliams Way	MODESTO	CA 95351	USA	(260) 205 1471			Shirley.J.Gonzalez@snomail.com	08/07/1
31	ESV455997	Mr William Thompson	745 Greenwood Rd	WEST COLUMBIA	SC 29169		(865) 043 3742	(571) 764 1889	(819) 417 7233	William.S.Thompson@xmail.com	08/15/1
32	FNR660911	Mrs Edith Reed	3619 North 35th Avenue	PHOENIX	AZ 85017	USA	(480) 896 5225				12/20/1

Record 61 of 100

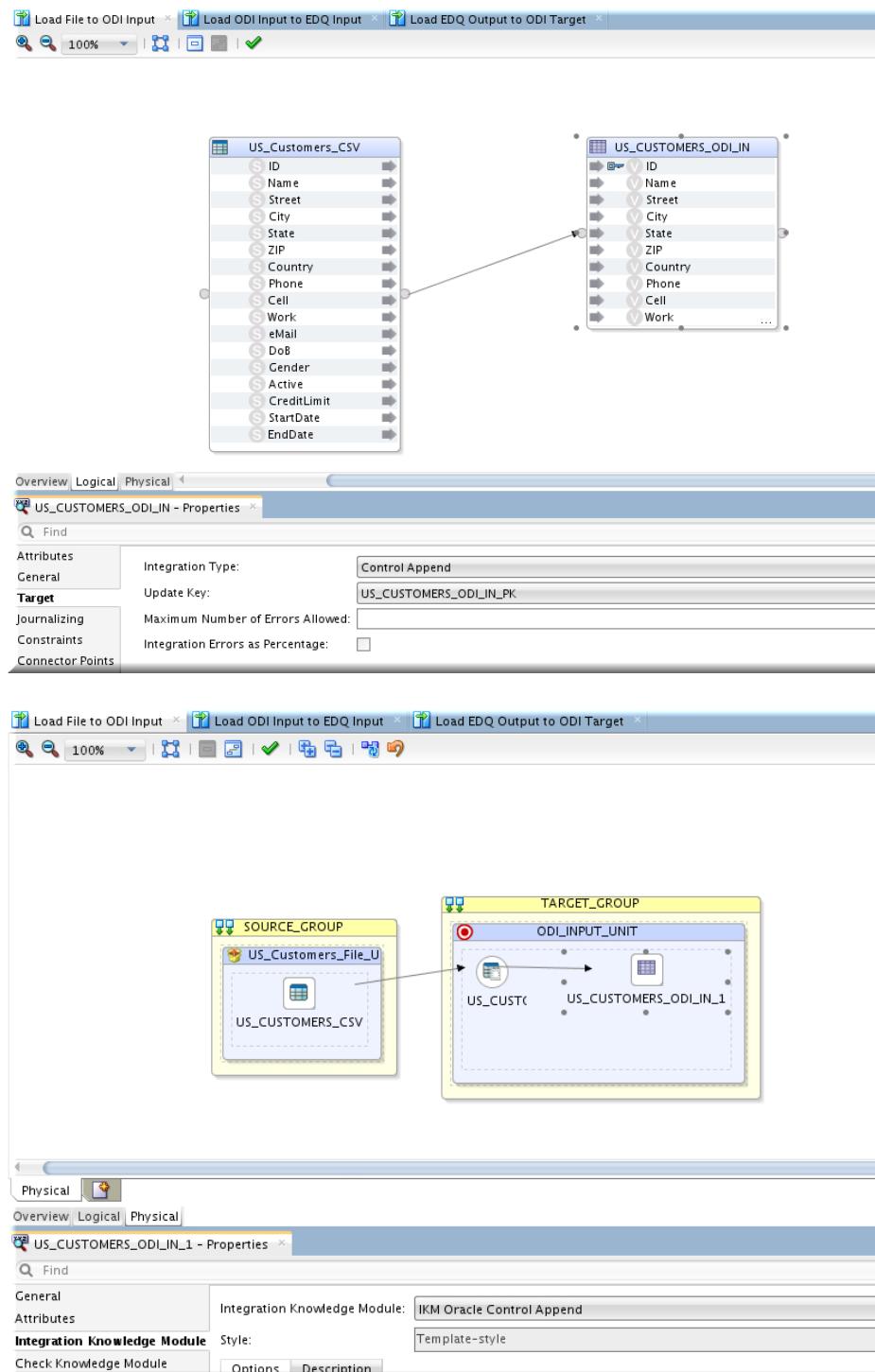
Tạo ánh xạ trong ODI

Ba ánh xạ sẽ được tạo ở phần tiếp theo để cho phép tải tệp vào các bảng cơ sở dữ liệu.

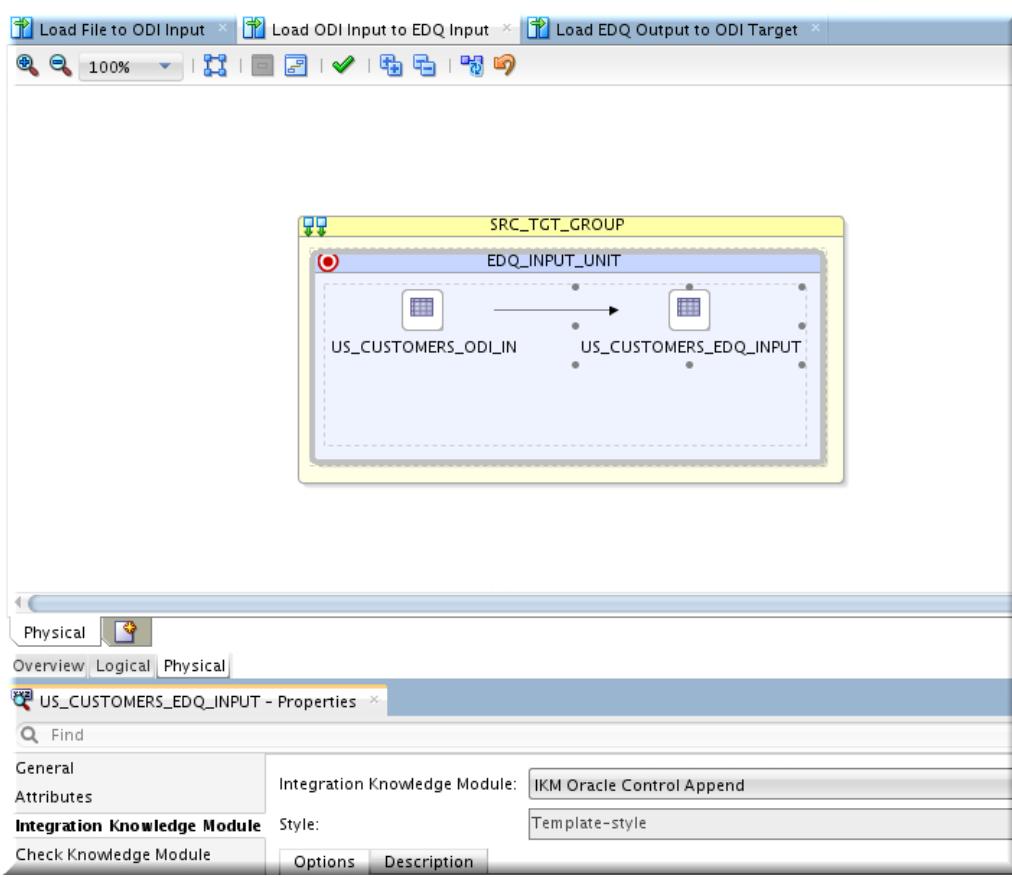
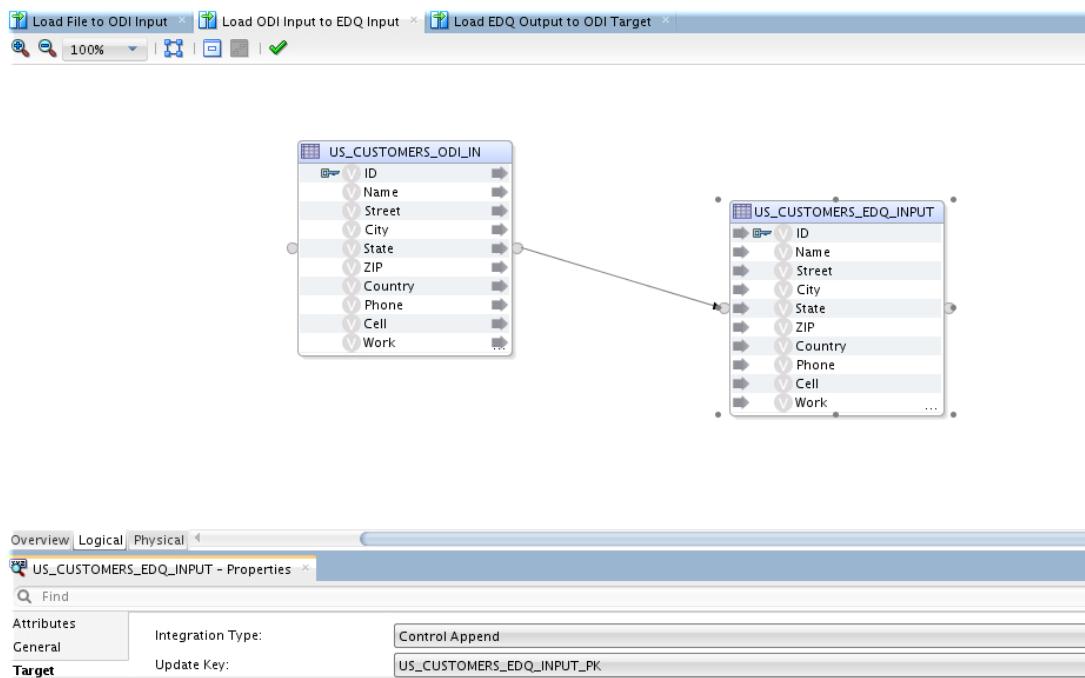
- Chuyển động tọa độ đầu tiên từ dữ liệu thô, một tệp giá trị được phân tách bằng dấu phẩy **Khách hàng Hoa Kỳ.csv** đến một bảng Oracle, **KHÁCH HÀNG US_ODI_IN**.
- Thứ hai sử dụng Đầu vào ODI, **KHÁCH HÀNG US_ODI_IN** như là nguồn để tải bảng Đầu vào **EDQUS_CUSTOMERS_EDQ_INPUT**.
- Cuối cùng sử dụng bảng Đầu ra EDQ **KHÁCH HÀNG US_ĐÃ_SẠCH_SẠCH** để tải bảng ODI Target **MỤC TIÊU CỦA KHÁCH HÀNG USODI**.
- Mỗi ánh xạ sẽ sử dụng kỹ thuật tích hợp Control Append.

Điều đáng chú ý là mặc dù không có ánh xạ trực tiếp trong ODI để tải bảng Đầu ra EDQ **KHÁCH HÀNG US_ĐÃ_SẠCH_SẠCH**, EDQ sẽ ghi dữ liệu vào bảng đó sau khi hoàn tất thành công tác kiểm tra chất lượng dữ liệu mà ODI sẽ gọi.

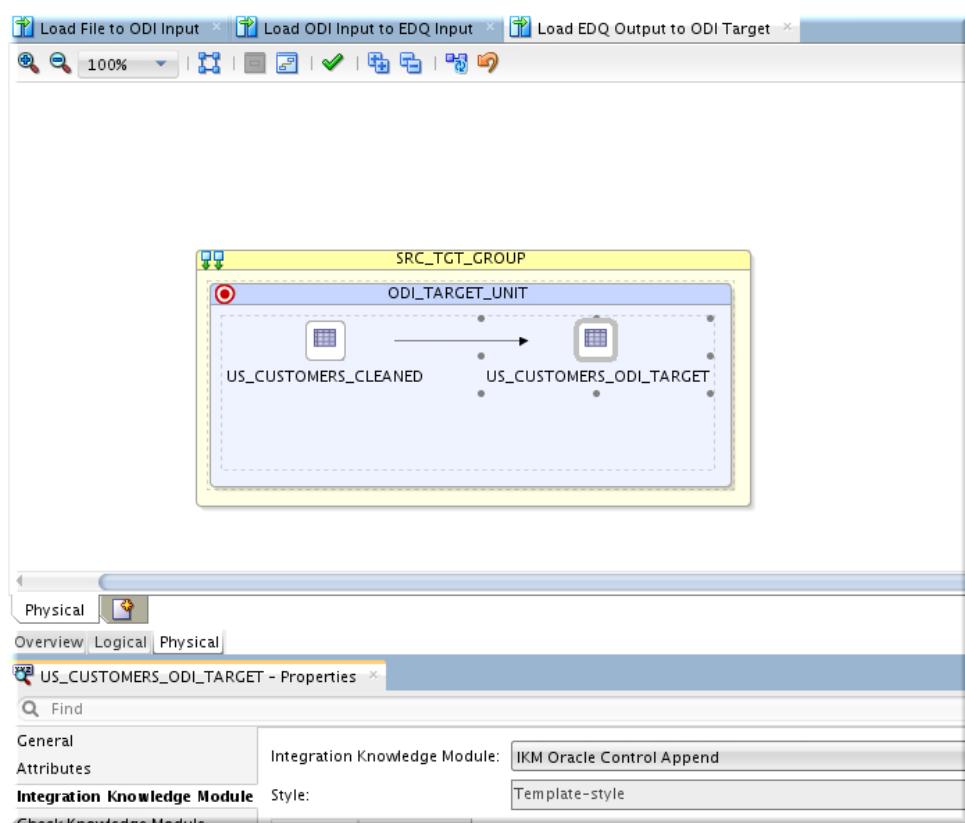
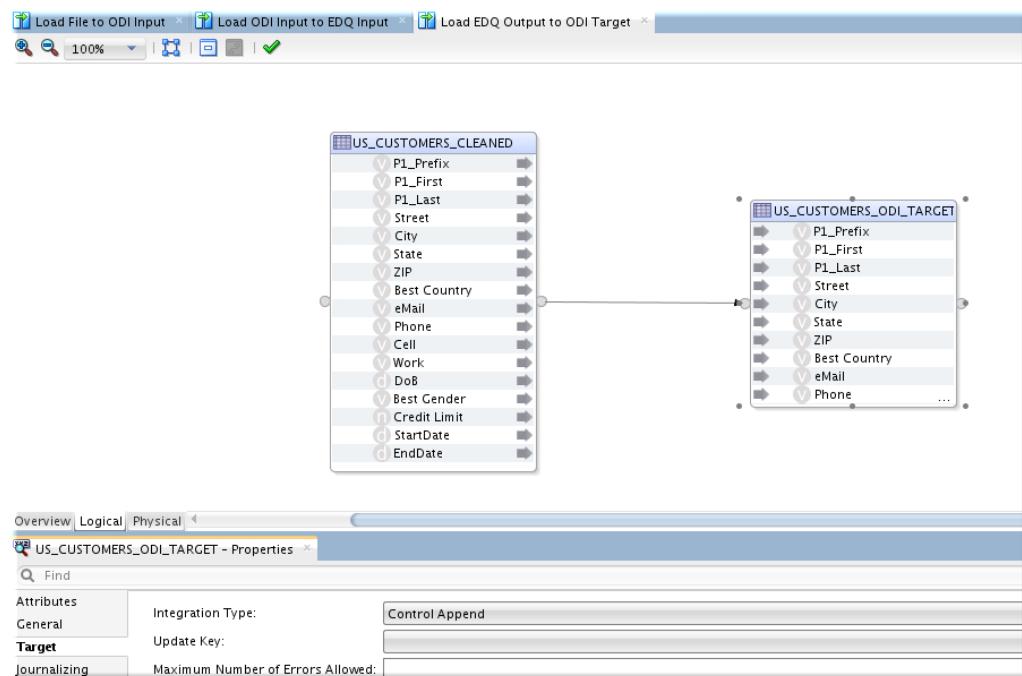
19. Mô hình **KHÁCH HÀNG US_ODI_IN** đóng vai trò là Đầu vào ODI cho EDQ và được tải từ **Khách hàng US** tập tin. Một bản đồ **Tải tệp vào đầu vào ODI** đã được tạo. Vì các cột giữa tệp và bảng giống hệt nhau trong ví dụ này nên tùy chọn tự động ánh xạ đã được sử dụng.



20. Một bản đồ khác **Tải đầu vào ODI vào đầu vào EDQ**được tạo ra như hình dưới đây



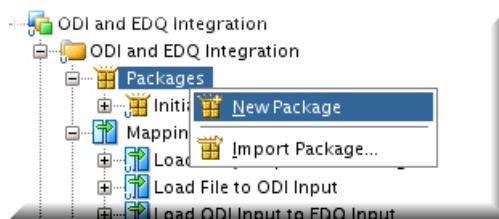
21. Cuối cùng, một bản đồ **Tải đầu ra EDQ vào mục tiêu ODI** được tạo ra như hình dưới đây



Tạo một gói ODI

Gói ODI sẽ đóng vai trò là phương tiện tự động hóa các tác vụ diễn ra trong quá trình thực hiện từng ánh xạ. Ngoài ra, Công cụ mở cho Chất lượng dữ liệu doanh nghiệp sẽ được sử dụng trong gói để gọi tác vụ EDQ hiện có.

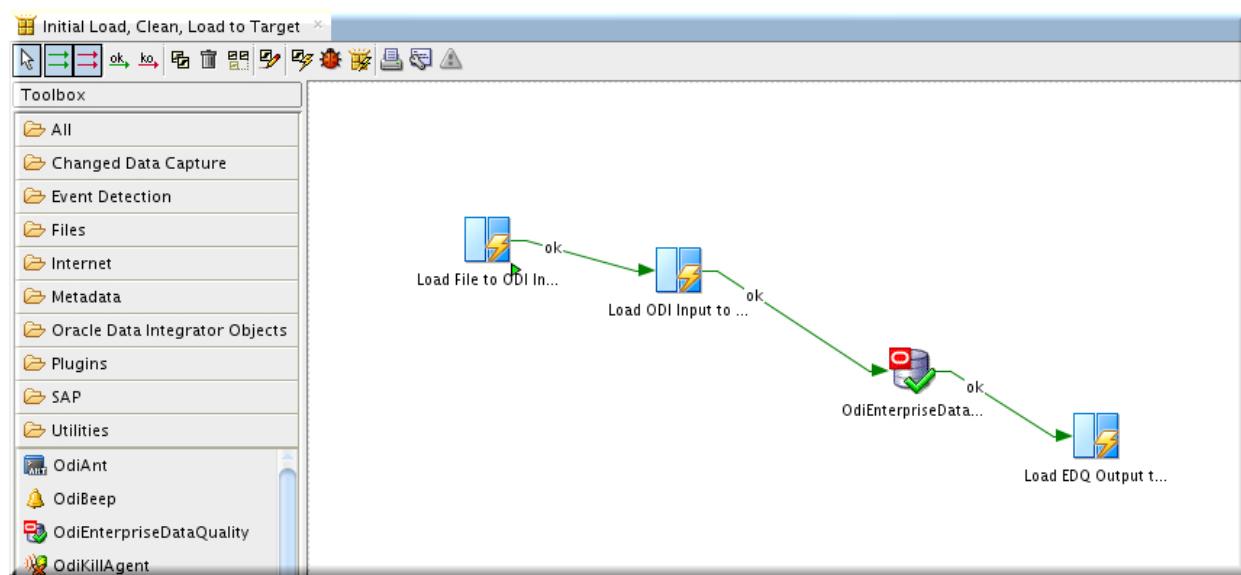
- Nhấp chuột phải vào **Gói hàng** trong vòng **Nhà thiết kế** tab và **Dự án** phần này tiết lộ tùy chọn để tạo một **Gói mới**



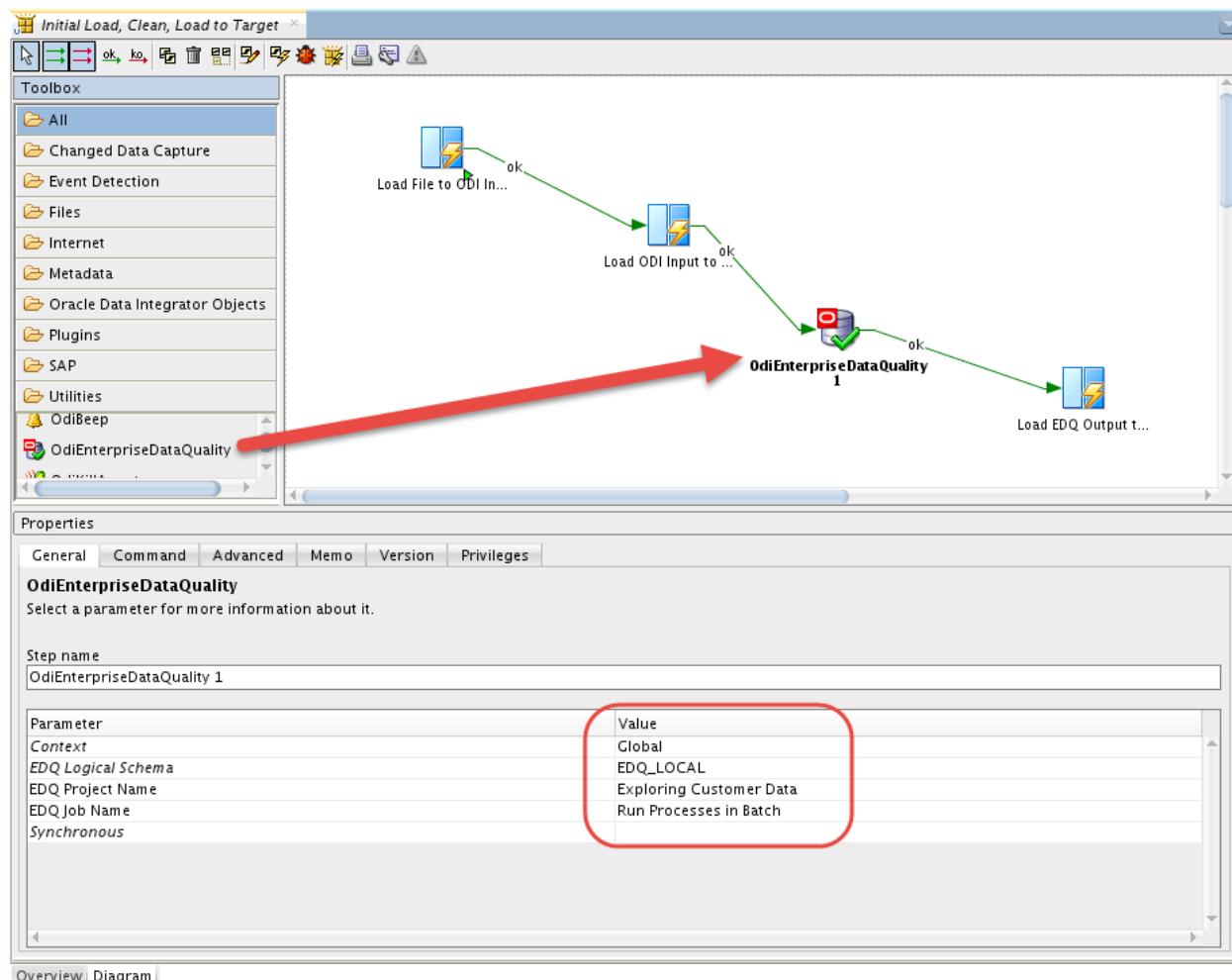
- Ba phép ánh xạ từ **Nhà thiết kế** tab được kéo và thả vào canvas trong **Bưu kiện** tab. Trong trường hợp này, gói được tạo ra được đặt tên là **Tải ban đầu, Làm sạch, Tải đến mục tiêu**

- Các **OdiEnterpriseDataChất lượng** Công cụ mở đã được kéo và thả vào canvas

- Cuối cùng **ĐƯỢC RỜI** tùy chọn đã được chọn để tạo một chuỗi để thực hiện các ánh xạ và gọi tác vụ EDQ khi hoàn tất thành công từng bước như trong ảnh chụp màn hình bên dưới



26. Nhấp vào **OdiEnterpriseDataChất lượng** biểu tượng trong canvas hiển thị cửa sổ thuộc tính ở phía dưới **Phòng thu ODI**

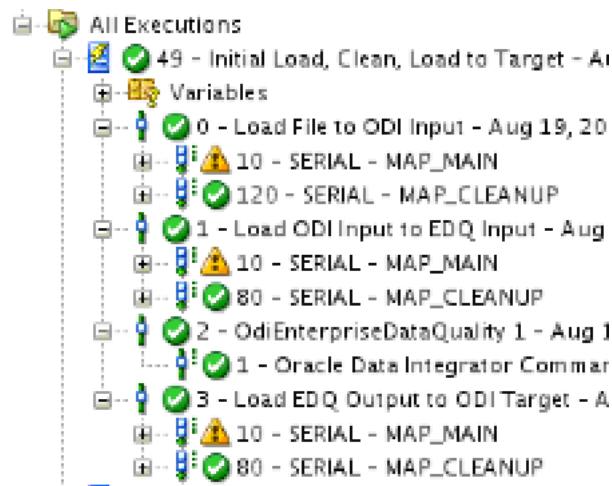


27. Có 6 **Các tham số** để điền vào để gọi thành công một công việc từ EDQ từ một Gói ODI. **Đồng bộ** tuy chọn sẽ sử dụng giá trị mặc định (Có), vì vậy có thể để trống

Bối cảnh	Bối cảnh thực thi ODI từ ODI Topology
Sơ đồ logic EDQ	Sơ đồ logic EDQ từ Topology ODI
Tên dự án EDQ	Tên dự án EDQ
Tên công việc EDQ	Tên công việc EDQ từ Dự án EDQ

28. Sau khi các thông số được nhập vào. Nhấp chuột phải vào **Bưu kiện** (Tải ban đầu, Làm sạch, Tải đến mục tiêu) để bắt đầu quá trình tải tệp .CSV vào các bảng phân đoạn. Vì OK- mũi tên được sử dụng cho mỗi bản đồ và công cụ mở, miễn là mỗi bản đồ là

thành công, mỗi bước sẽ tiếp tục. Việc thực hiện có thể được quan sát trong **Người điều hành tab**



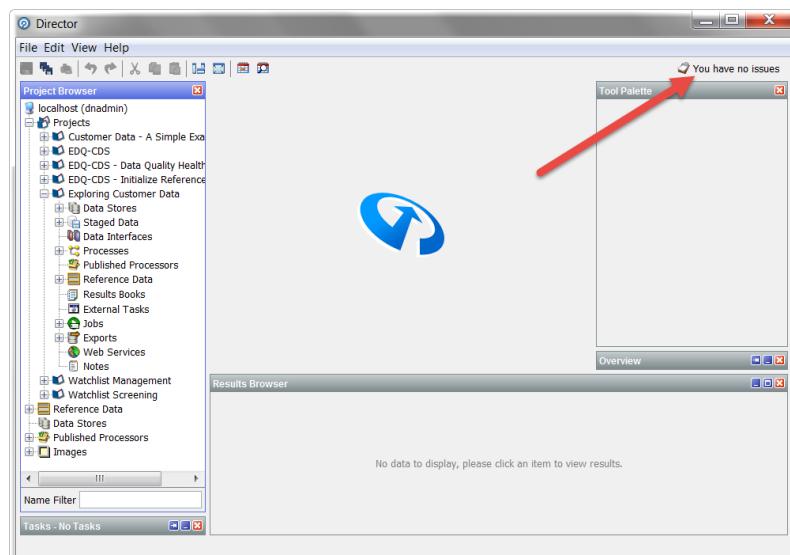
29. Quay trở lại **Nhà thiết kế tab** và **Mô hình phần**. Nhấp chuột phải vào ODI Target **MỤC TIÊU CỦA KHÁCH HÀNG USODI** cho phép bạn xem dữ liệu đã được làm sạch, tiêu đề cột **Lưu ý** tại siêu dữ liệu được tạo bởi Enterprise Data Quality. cột cho quốc gia được chuẩn hóa **Ngoài ra**, một tập hợp các giá trị thống nhất

P1_Prefix	P1_First	P1_Last	Street	City	State	ZIP	Best Country	eMail	Phone	Cell	Work
1 Mr	Bub	Mayo	2100 North Lark Drive	FENTON	MO	63026	United States of America	Robert.C.Mayo@shockmail.com	(238) 777 8897		
2 Mr	Brian	Robles	10800 Foreman	LOWELL	MI	49330	United States of America	Brian.K.Robles@xtmail.com	(740) 937 3714		
3 Mr	Perry	Elbert	202 C Street	SAN DIEGO	CA	92101	United States of America		(644) 6935745		
4 Mr	Walter	Torrell	442 East 144th Avenue	BROOMFIELD	CO	80020	United States of America		(330) 738 2431	(589) 048 1176	
5 Mr	Henry	Segovia	1 Walsh Drive	PARAGOULD	AR	72450	United States of America		do not call		
6 Mrs	Angela	Szymanski	1450 Jeffco Boulevard	ARNOLD	MO	63010	United States of America	Angela.G.Szymanski@scene46.com	(532) 137 1134		
7 Mr	Arthur	Thompson	455 Grand Bay Drive	KEY BISCAYNE	FL	33149	United States of America		(578) 861 3129	(809) 372 6839	
8 Mr	J	Turner	123 W.C. Acker	PICKENS	SC	29671	United States of America		(880) 436 8398		
9 Mrs	Elizabeth	Wenzel	4011 Blue Ridge Cut Off	KANSAS CITY	MO	64133	United States of America	Elizabeth.C.Wenzel@shockmail.com	(892) 548 2681	(585) 229 6224	
10 Mr	Ray	Machado	2120 W Guadalupe Rd	GILBERT	AZ	85233	United States of America	Ray.T.Machado@xtmail.com	(146) 947 2347		
11 Mrs	Sheryl	Miller	400 South Highland	JACKSON	TN	38301	United States of America		(665) 174 2747	(134) 486 1452	
12 Mrs	Jessica	Martinez	10094 Premier Parkway	MIRAMAR	FL	33425	United States of America		do not call		
13 Mrs	Kimberly	Collins	200 N 5th Street	GARLAND	TX	75040	United States of America		(162) 005 4134		
14 Mr	Henry	Rankin	13181 Hanover Courthouse Road	HANOVER	VA	23069	United States of America	Henry.C.Rankin@shockmail.com	(303) 808 4391		
15 Mr	Fred	Martin	1701 Sharp Road	WATERFORD	WI	53185	United States of America	Fred.G.Martin@hu.com	(309) 596 9871		
16 Mr	Young	Ryan	700 W 20 Street	HIALEAH	FL	33010	United States of America	Young.T.Ryan@scene46.com	(880) 272 7345	(400) 973 7608	(417) 253 2
17 Mrs	Martina	Smooth	159 Dwight Park Circle	SYRACUSE	NY	13209	United States of America		(589) 766 0848		
18 Mrs	Ide	Welker	600 Main Street	JOHNSON CITY	NY	13790	United States of America	Ida.M.Welker@snowmail.com	(339) 823 6735	(894) 663 3399	
19 Mrs	Catherine	Taylor	9103 E 39th St	KANSAS CITY	MO	64133	United States of America	Catherine.A.Taylor@xtmail.com	(443) 966 1586	(446) 559 4404	(562) 147 9
20 Mr	Steven	Wolfe	133 Freeport Road	PITTSBURGH	PA	15215	United States of America		(293) 424 5675		
21 Mr	Brett	Ellison	1270 South Lipan Street	DENVER	CO	80223	United States of America		(285) 596 1893		
22 Mrs	Marla	Summer	3614 Security Street	GARLAND	TX	75040	United States of America	Marla.T.Summer@scene46.com	(551) 365 3320		
23 Mrs	Kim	Said	305 College Street NE	LACEY	WA	98516	United States of America	Kim.J.Said@xtmail.com	(333) 642 4202		
24 Mrs	Anna	Jones	1533 E Lindsay Street	STOCKTON	CA	95205	United States of America		(190) 512 9345		
25 Mr	Dennis	Taylor	505 Park Avenue	ODESSA	TX	79761	United States of America	Dennis.M.Taylor@xtmail.com	(845) 727 9800	(340) 920 1471	
26 Mrs	Vicki	Samuels	4559 Peoples Road	PITTSBURGH	PA	15237	United States of America		(436) 595 4591	(647) 288 0218	
27 Mrs	Stephanie	Cox	20656 84th Avenue South	KENT	WA	98032	United States of America	Stephanie.W.Cox@scene46.com	(118) 413 3256		
28 Mrs	Kimberly	Hardison	108 Miller Avenue	JACKSON	TN	38305	United States of America	Kimberly.C.Hardison@xtmail.com	(776) 766 4272		
29 Mr	Edward	Brooker	549 E Texas Street	VAN	TX	75790	United States of America		(357) 585 5285	(464) 628 4827	
30 Mrs	Aimee	Valdez	14260 SW 136 Street Bay 20	MIAMI	FL	33186	United States of America		(117) 005 6304		
31 Mrs	Deanna	Olipphant	3815 River Crossing Parkway	INDIANAPOLIS	IN	46240	United States of America	Deanna.A.Olipphant@hu.com	(642) 283 6511		
32 Mr	Timothy	Daniel	700 Haywood Road	GREENVILLE	SC	29607	United States of America	Timothy.V.Daniel@scene46.com	(125) 924 4126	(743) 118 3771	(594) 805 7

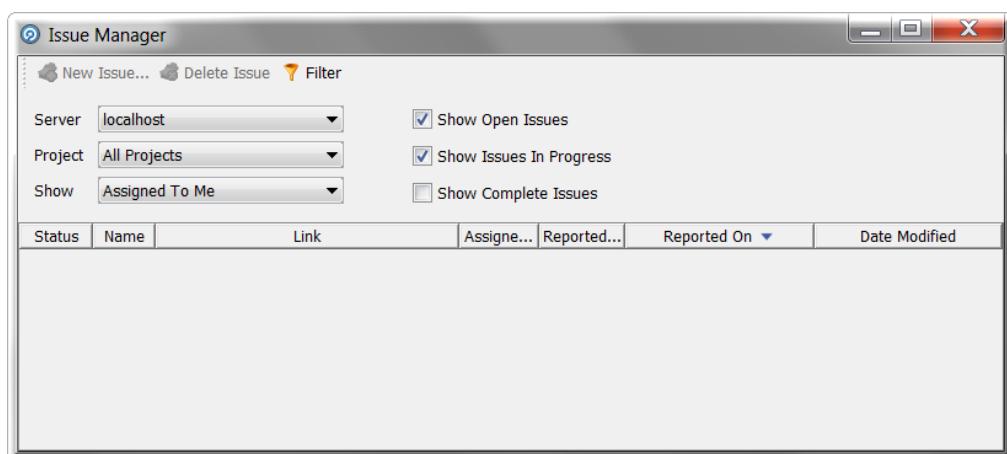
Phòng thí nghiệm 6: Quản lý vấn đề

Issue Management trong EDQ cho phép bạn có phương tiện để theo dõi các vấn đề về chất lượng dữ liệu cần được giải quyết. Các vấn đề có thể được chỉ định cho bạn, người dùng khác hoặc một nhóm để điều tra và giải quyết. Có thể tìm thấy bảng điều khiển Issue Manager trong ứng dụng Director hoặc thông qua EDQ Launchpad

- Điều hướng đến **Quản lý vấn đề** trong ứng dụng EDQ Director. Tìm biểu tượng ở góc trên bên phải của Director. Tùy thuộc vào người bạn giao các vấn đề được tạo trong phòng thí nghiệm trước đó, có thể có một **tin nhắn cho biết bạn đang gặp một số vấn đề đang chờ**



- Hãy dành chút thời gian để khám phá giao diện của **Quản lý vấn đề**





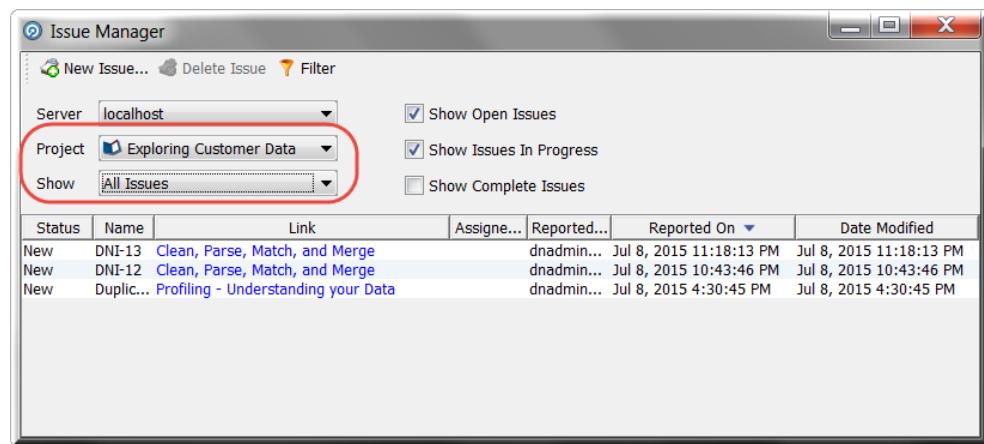
Issue Manager vừa được mở bằng ứng dụng Director, nhưng đối với người dùng chỉ chịu trách nhiệm quản lý vấn đề, ứng dụng cũng có thể được mở thông qua EDQ Launchpad (có thể truy cập tại <http://localhost:8011/edq/faces/> trong phòng thí nghiệm này)

The screenshot shows the Oracle Enterprise Data Quality Launchpad. At the top, there is a message: "Issue Manager vừa được mở bằng ứng dụng Director, nhưng đối với người dùng chỉ chịu trách nhiệm quản lý vấn đề, ứng dụng cũng có thể được mở thông qua EDQ Launchpad (có thể truy cập tại <http://localhost:8011/edq/faces/> trong phòng thí nghiệm này)". Below this, the Launchpad interface is displayed with several tiles:

- Director
- Server Console
- Match Review
- Case Management
- Case Management Administration
- Configuration Analysis
- Issue Manager** (this tile has a red arrow pointing to it)
- Watchlist Screening

At the bottom of the page, there are links for Home, Privacy Statement, and About, followed by a copyright notice: "Copyright © 2006, 2015, Oracle and/or its affiliates. All rights reserved."

3. Trong Trình quản lý sự cố, hãy bắt đầu bằng cách thay đổi danh sách thả xuống cho **Dự án** **ĐẾN** **Khám phá dữ liệu khách hàng** và **Trình diễn** **ĐẾN** **Tất cả các vấn đề**

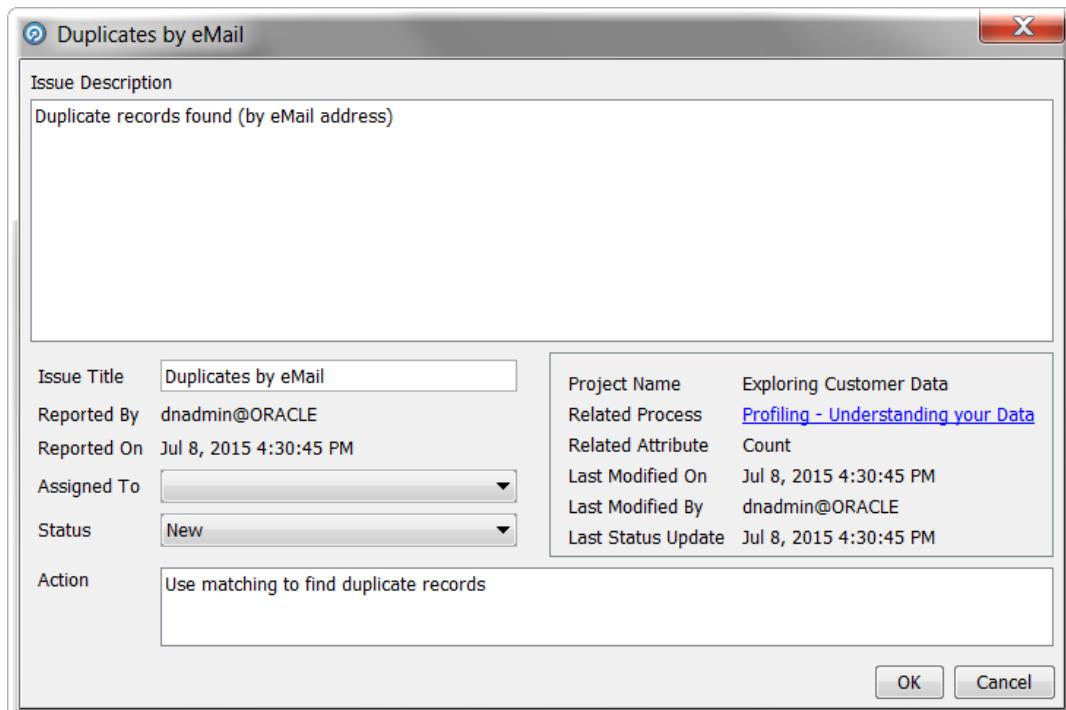


Trình quản lý sự cố sẽ hiển thị các sự cố theo Trạng thái, Tên, Liên kết (nơi sự cố được mở), Nhiệm vụ, Người dùng đã mở sự cố, Ngày báo cáo và lần cuối cùng sự cố được sửa đổi.

4. Mở một vấn đề bằng cách nhấp đúp vào bản ghi bên dưới **Tên**cột trong **Quản lý vấn đề**

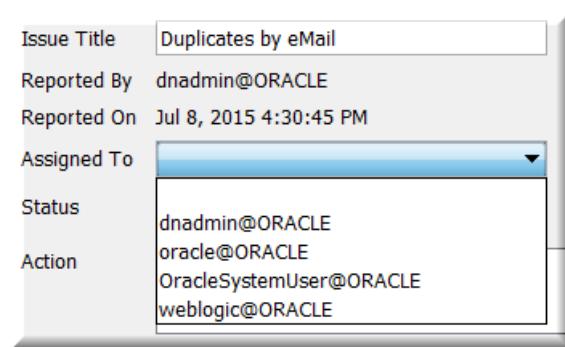


Các **Liên kết**cột là một siêu liên kết đang hoạt động sẽ đưa ứng dụng Director trở lại vị trí mà sự cố đã được mở. Bạn cũng có thể mở một sự cố bằng cách nhấp chuột phải vào một bản ghi và nhấp vào **Vấn đề mở**.

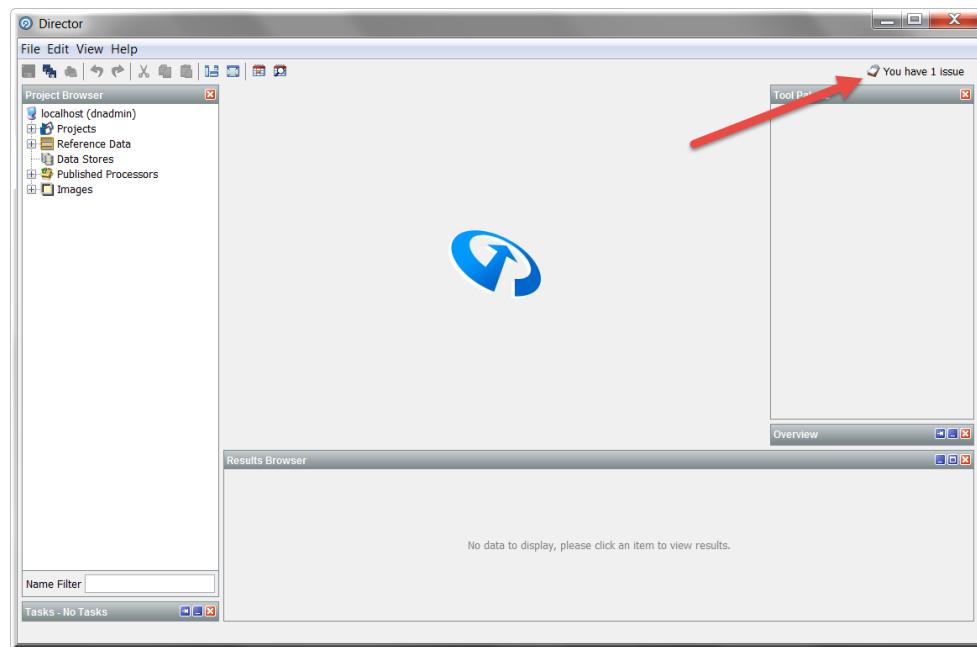


Các **Tiêu đề vấn đề** phản ánh cột Tên từ Trình quản lý sự cố. Ngoài ra, có thể ghi chú một mục hành động ở cuối cho những cá nhân được giao nhiệm vụ giải quyết sự cố. Chi tiết bổ sung được cung cấp ở góc dưới bên phải của cửa sổ sự cố. Lưu ý rằng Tên dự án, Quy trình liên quan và Thuộc tính liên quan được ghi lại.

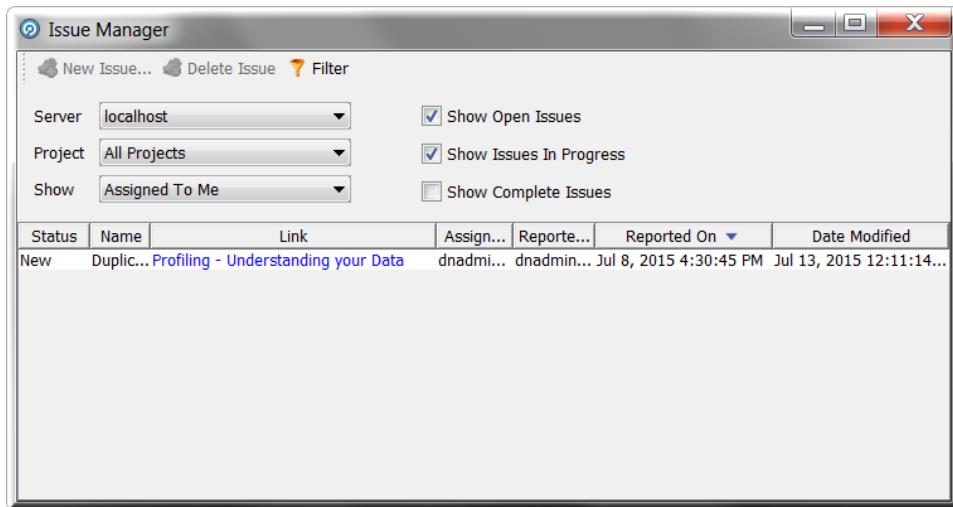
- Chọn danh sách thả xuống có tiêu đề **Được giao cho** và giao vấn đề chodnadmin@ORACLE, nhấp chuột **ĐƯỢC RỒI** để tiếp tục



Tại thời điểm này, bạn sẽ có ít nhất 1 vấn đề được chỉ định cho người dùng mà bạn đã đăng nhập. Lưu ý **góc trên bên phải** của **Giám đốc** phản ánh điều này.

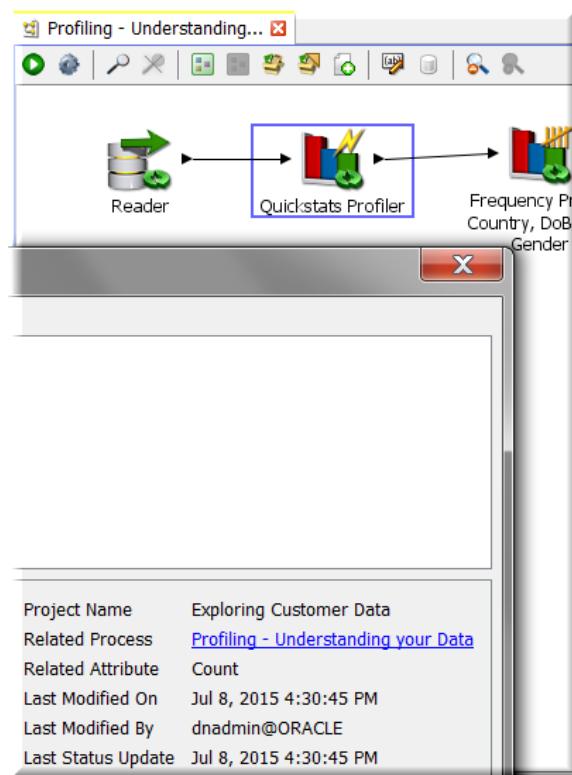


6. Quay trở lại **Quản lý vấn đề** bằng cách nhấp vào biểu tượng hoặc thông báo cho biết bạn có vấn đề. Lưu ý theo mặc định các bộ lọc cho **Dự án** và **Trình diễn** được mặc định là **Tất cả các dự án** và **Giao cho tôi**. Điều này đảm bảo rằng khi bạn đăng nhập, bạn sẽ thấy tất cả các vấn đề quan trọng mà bạn được giao để giải quyết



Nếu bạn nhấp vào **Lọc** ở đầu Trình quản lý sự cố, bạn sẽ thấy menu thả xuống và các tùy chọn khác ở phía bên phải của Trình quản lý sự cố bị ẩn.

7. Nhấp vào siêu liên kết bên dưới **Liên kết cột**, và quay lại Giám đốc. Bạn sẽ nhận thấy rằng nó đưa bạn đến dự án của bạn và **Phân tích dữ liệu - Hiểu dữ liệu của bạn** quá trình. Hơn nữa, lưu ý rằng **Hồ sơ Quickstats** đã được chọn vì đó là nơi vấn đề được tạo ra



8. Các **Trình duyệt kết quả** có thể nêu 'Quy trình này cần được chạy trước khi xem kết quả'. Nhấp vào nút chạy ở góc trên bên trái của cửa sổ Director để chạy quy trình. Sau đó, người dùng có thể điều tra thêm vấn đề bằng cách đi sâu vào các kết quả tìm thấy trong Trình duyệt kết quả

Khi người dùng bắt đầu điều tra và khắc phục các vấn đề về chất lượng dữ liệu được nêu ra bằng chức năng Trình quản lý sự cố, điều quan trọng là phải thay đổi **Trạng thái ĐỀNĐang tiến hành** hoặc **Hoàn thành** để người dùng EDQ có thể tiếp tục giải quyết các vấn đề mới hoặc chưa được phân công một cách hiệu quả.

Tính năng Quản lý sự cố của EDQ cho phép người dùng của bạn cộng tác khi làm việc trên các sáng kiến và dự án về Chất lượng dữ liệu. Ngoài chức năng có trong Director hoặc Issue Manager, EDQ cũng có thể được cấu hình để thông báo cho người dùng khi có sự cố mới phát sinh.

Mẹo và thủ thuật triển khai và sử dụng Oracle Data Integrator và EDQ cùng nhau

Trong phần này, bạn sẽ xem xét một số phương pháp hay nhất liên quan đến việc sử dụng và triển khai Enterprise Data Quality. Ngoài các phương pháp hay nhất, bạn có thể tìm thấy danh sách ngắn các tài nguyên bên dưới.

- EDQ có nhiều thuộc tính có thể được sửa đổi để tinh chỉnh môi trường. Có thể tìm thấy hướng dẫn đầy đủ [đây](#)
ôi http://docs.oracle.com/middleware/1213/edq/DQSAG/perf_tuning.htm

Nhiều điều khiển điều chỉnh được giải thích chi tiết trong tài liệu được liên kết ở trên, nhưng hai điều khiển phổ biến nhất được thảo luận bên dưới:

- Một trong những điều chỉnh phổ biến mà nhiều môi trường thực hiện là sửa đổi Kích thước Heap của máy khách do các sự cố có thể xảy ra khi bạn cố gắng xuất một lượng lớn dữ liệu sang tệp Excel phía máy khách.

- Ôi Tìm thấy bản thiết kế.properties tập tin trong cấu hình/thuộc tính thư mục của máy chủ EDQ
- Ôi Bạn sẽ thấy *.jvm.memory = 512m trong tập tin đã có. Bộ nhớ phân bổ có thể được thay đổi để ảnh hưởng đến tất cả các ứng dụng khách hàng (tức là Giám đốc, Quản lý sự cố, v.v.). Để sửa đổi kích thước đồng cho một ứng dụng khách hàng nhất định, hãy thay thế * với giá trị để thay đổi nó cho Giám đốc hoặc vấn đề để thay đổi nó cho Issue Manager
- Ôi Danh sách đầy đủ các ứng dụng của khách hàng có thể được tìm thấy trong liên kết ở trên để tinh chỉnh môi trường

- EDQ có nhu cầu sử dụng không gian đĩa rất lớn trong thời gian chạy. Hãy đảm bảo thiết lập kích thước không gian bằng cơ sở dữ liệu của bạn một cách phù hợp để tính đến hệ số lạm phát kích thước dữ liệu từ 10x đến 25x. Vì vậy, nếu dữ liệu nguồn của bạn là 2GB, bạn sẽ cần từ 20 đến 50GB không gian bằng khi chạy quy trình EDQ của mình (đặc biệt nếu một Quy trình nhất định đang sử dụng quy trình Profiler)

- Trong trường hợp máy chủ cơ sở dữ liệu mạnh hơn máy chủ lưu trữ ứng dụng EDQ, có thể đáng để tăng workunitexecutor.outputThreads thành số để sửa đổi số luồng (và kết nối cơ sở dữ liệu) được sử dụng khi ghi kết quả và dữ liệu được dàn dựng vào cơ sở dữ liệu
 - Ôi Tham số này được tìm thấy trong `giám đốc.properties` tập tin trong Trang chủ thư mục cấu hình

Tài nguyên bổ sung

- [Thư viện học tập về chất lượng dữ liệu doanh nghiệp Oracle](#)
 - Ôi <http://oracle.com/oll> - Nhấp vào Tìm kiếm và sử dụng menu thả xuống để điều chỉnh Bộ lọc tìm kiếm thành:
 - Gia đình sản phẩm – Fusion Middleware
 - Sản phẩm – Chất lượng dữ liệu doanh nghiệp
- [Tài liệu Oracle – Oracle Enterprise Data Quality 12.2.1](#)
 - Ôi <http://docs.oracle.com/middleware/1221/edq/index.html>
- [Trang chủ Oracle Enterprise Data Quality](#)
 - Ôi <http://www.oracle.com/technetwork/middleware/oedq/overview/index.html>
- [Blog tích hợp dữ liệu Oracle](#)
 - Ôi <https://blogs.oracle.com/dataintegration/>