TRƯỜNG ĐẠI HỌC KINH TẾ

**KHOA THỐNG KÊ – TIN HỌC**

A blue and white logo

Description automatically generated

**BÁO CÁO THỰC TẬP NGHỀ NGHIỆP**

**NGÀNH HỆ THỐNG THÔNG TIN QUẢN LÝ**

**CHUYÊN NGÀNH QUẢN TRỊ HỆ THỐNG THÔNG TIN**

**PROCESS AND ANALYZE DATA WITH THE CLOUD**

Sinh viên thực hiện : **Dương Thị Kiều Doan**

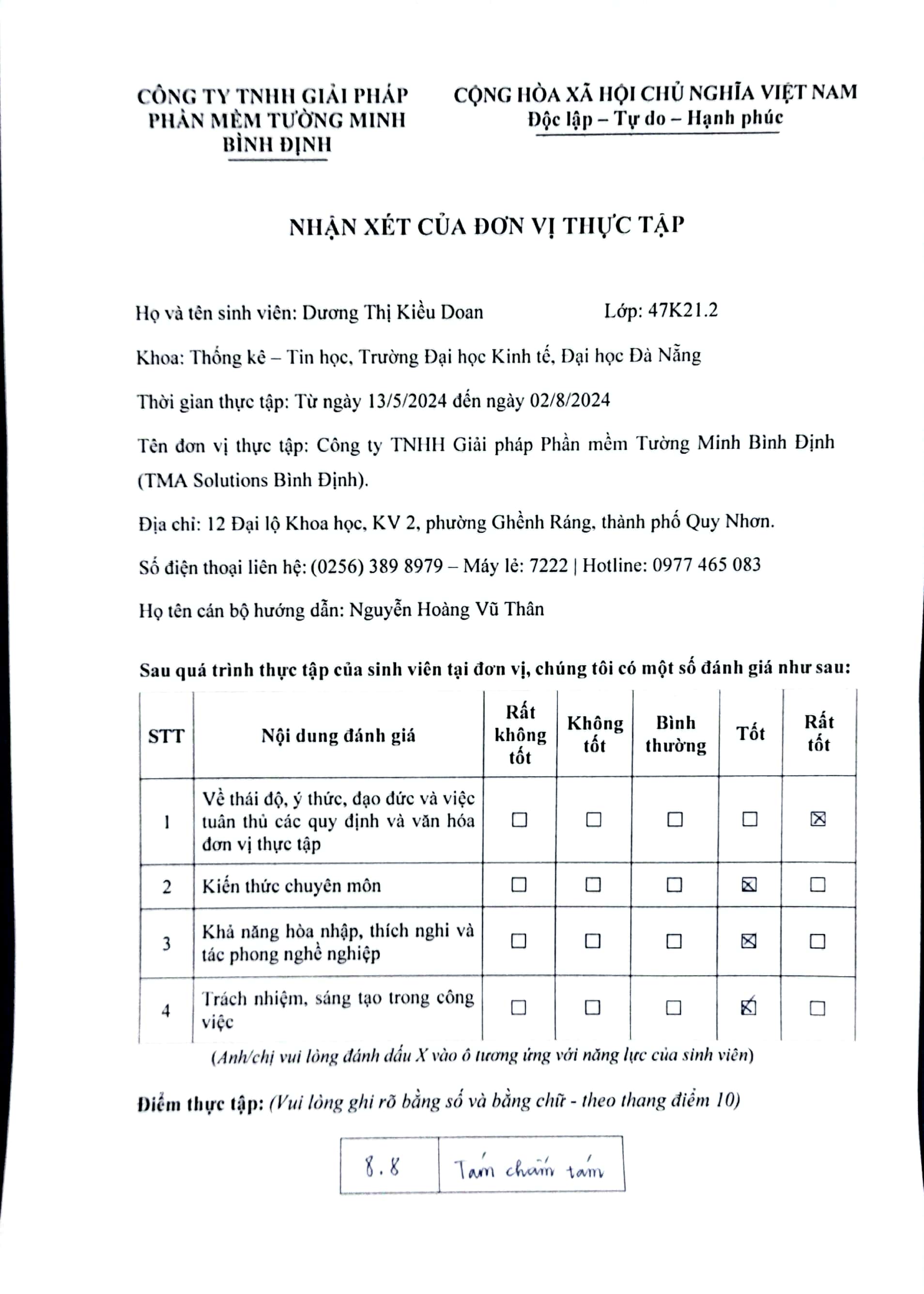
Lớp : **47K21.2**

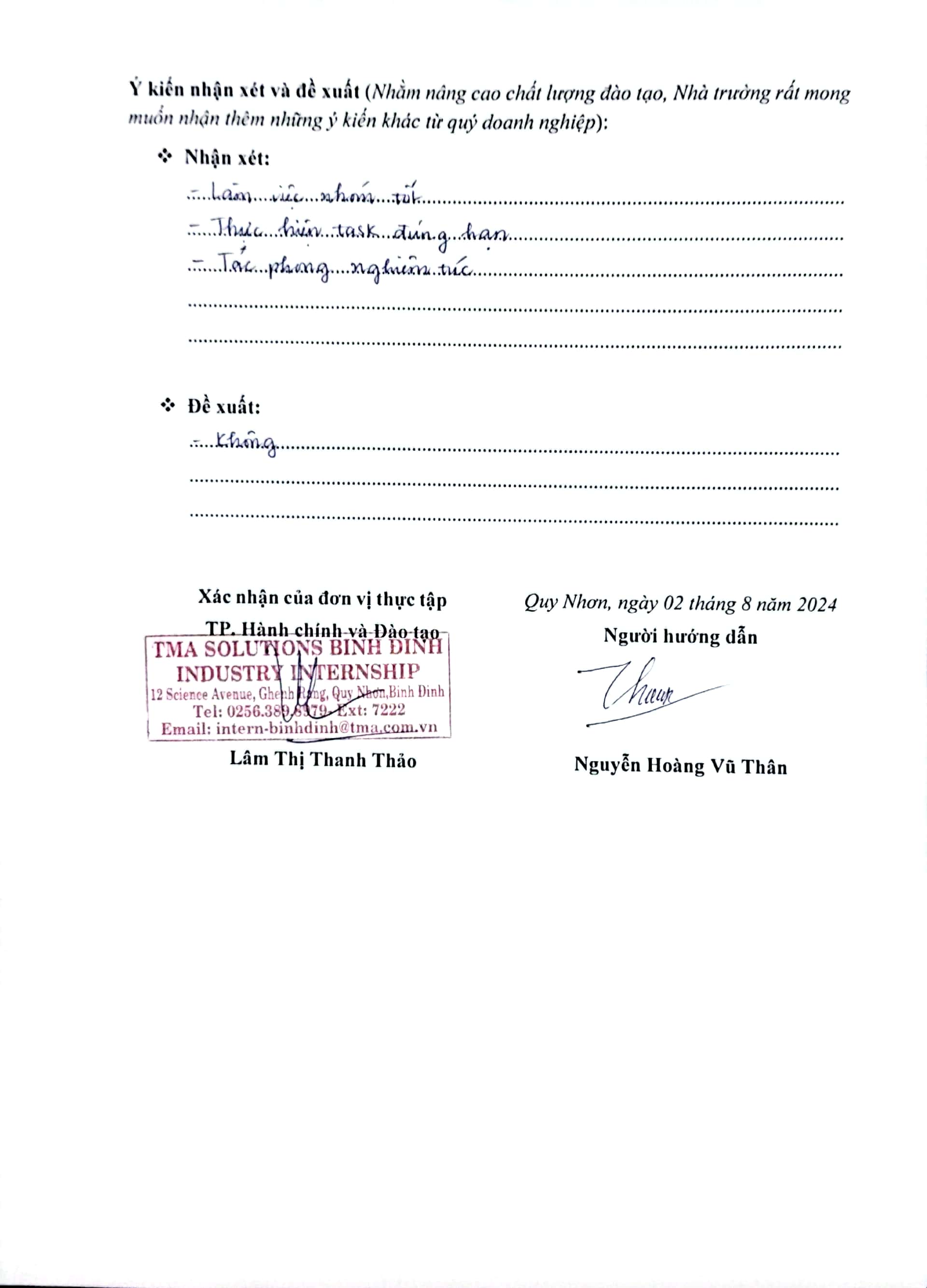
Đơn vị thực tập : **TMA Solutions Bình Định**

Cán bộ hướng dẫn : **Nguyễn Hoàng Vũ Thân**

Giảng viên hướng dẫn : **TS. Hoàng Thị Thanh Hà**

***Đà Nẵng, 8/20******24***

**NHẬN XÉT CỦA ĐƠN VỊ THỰC TẬP**  




# LỜI CẢM ƠN

Lời đầu tiên, em xin chân thành cảm ơn đến quý thầy cô trường Đại học Kinh Tế nói chung và quý thầy cô khoa Thống Kê - Tin học nói riêng đã giúp em nắm vững kiến thức cơ sở trong quá trình học tập tại trường. Đặc biệt hơn, em gửi lời cảm ơn sâu sắc đến cô Hoàng Thị Thanh Hà, người đã hỗ trợ, quan tâm, tận tình hướng dẫn để em có thể hoàn thành tốt nhất bài báo cáo thực tập kỳ hè này.

Em gửi lời cảm ơn chân thành nhất đến Ban Lãnh đạo và toàn thể các anh chị trong công ty TMA Solutions Bình Định đã tạo điều kiện cho em có cơ hội được học tập và làm việc trong một môi trường chuyên nghiệp của công ty. Đặc biệt, em vô cùng biết ơn đến anh Nguyễn Hoàng Vũ Thân và chị Tăng Thị Thúy Vân đã hướng dẫn, chia sẻ những kinh nghiệm quý báu trong suốt quá trình thực tập. Sự hỗ trợ này đã giúp em phát triển và hoàn thiện bản thân ngày một tốt hơn. Những kiến thức, kinh nghiệm ấy sẽ là hành trang để em có nền tảng cho sự nghiệp của bản thân trong tương lai.

Em xin chân thành cảm ơn ạ!

# LỜI CAM ĐOAN

Tôi xin cam đoan:

Nội dung và các tài liệu tôi cung cấp trong báo cáo này hoàn toàn là chính xác và được thực hiện hoàn thành dưới sự hướng dẫn của cô Hoàng Thị Thanh Hà.

Các tham khảo dùng trong Báo cáo đều được trích dẫn rõ ràng tên tác giả, tên công trình, thời gian, địa điểm công bố.

Tôi hiểu rằng, nếu có bất cứ thông tin nào không đúng sự thật hay có sự sao chép không hợp lệ, tôi sẽ chịu hoàn toàn trách nhiệm.

Sinh viên thực hiện

*Dương Thị Kiều Doan*

# MỤC LỤC

[LỜI CẢM ƠN iii](#_Toc173662638)

[LỜI CAM ĐOAN iv](#_Toc173662639)

[MỤC LỤC v](#_Toc173662640)

[DANH MỤC HÌNH ẢNH viii](#_Toc173662641)

[DANH MỤC BẢNG BIỂU x](#_Toc173662642)

[DANH MỤC CÁC TỪ VIẾT TẮT xi](#_Toc173662643)

[LỜI MỞ ĐẦU 1](#_Toc173662644)

[CHƯƠNG 1. TỔNG QUAN VỀ DOANH NGHIỆP 3](#_Toc173662645)

[1.1. Giới thiệu tổng quát về doanh nghiệp thực tập 3](#_Toc173662646)

[1.1.1. Thông tin chung về TMA Solutions Bình Định. 3](#_Toc173662647)

[1.1.2. Lĩnh vực hoạt động 4](#_Toc173662648)

[1.2. Tổng quan về vị trí việc làm 4](#_Toc173662649)

[1.2.1. Định nghĩa về Data Engineer 4](#_Toc173662650)

[1.2.2. Nhiệm vụ của một Data Engineer 4](#_Toc173662651)

[1.3. Các kỹ năng cần có để trở thành một Data Engineer 5](#_Toc173662652)

[1.3.1. Ngôn ngữ lập trình 5](#_Toc173662653)

[1.3.2. Cơ sở dữ liệu 5](#_Toc173662654)

[1.3.3. Big Data Technologies 5](#_Toc173662655)

[1.3.4. ETL (Extract, Transform, Load) 6](#_Toc173662656)

[1.3.5. Cloud Platforms 6](#_Toc173662657)

[1.3.6. Data Warehousing 6](#_Toc173662658)

[1.4. Cơ hội nghề nghiệp của Data Engineer 6](#_Toc173662659)

[1.5. Kết luận 7](#_Toc173662660)

[CHƯƠNG 2. CƠ SỞ LÝ THUYẾT VỀ DATA ENGINEER 8](#_Toc173662661)

[2.1. Giới thiệu về Data Engineering 8](#_Toc173662662)

[2.2. Các khái niệm cơ bản 8](#_Toc173662663)

[2.3. Lý thuyết liên quan 9](#_Toc173662664)

[2.4. Công nghệ và công cụ 9](#_Toc173662665)

[2.5. Kết luận 10](#_Toc173662666)

[CHƯƠNG 3. PHÂN TÍCH THIẾT KẾ DATABASE 11](#_Toc173662667)

[3.1. Khảo sát yêu cầu và phân tích 11](#_Toc173662668)

[3.2. Tổng quan về dữ liệu 11](#_Toc173662669)

[3.2.1. Giới thiệu tổng quan về dữ liệu 11](#_Toc173662670)

[3.2.2. Các bảng của dữ liệu 11](#_Toc173662671)

[a. Bảng Books 11](#_Toc173662672)

[b. Bảng Ratings 12](#_Toc173662673)

[c. Bảng Users 13](#_Toc173662674)

[3.2.3. Mối quan hệ của các bảng trong cơ sở dữ liệu 13](#_Toc173662675)

[3.3. Các công cụ cần dùng 14](#_Toc173662676)

[3.3.1. DataBricks (Pyspark) 14](#_Toc173662677)

[3.3.2. Azure Blob Storage 14](#_Toc173662678)

[3.3.3. Power BI 14](#_Toc173662679)

[3.4. Workflow của dự án 15](#_Toc173662680)

[3.4.1. Xử lý dữ liệu thô bằng DataBricks 15](#_Toc173662681)

[a. Đọc dữ liệu vào dataframe 15](#_Toc173662682)

[b. Xóa cột dữ liệu không sử dụng để phân tích 17](#_Toc173662683)

[c. Xóa dữ liệu trùng từ các cột khóa chính 17](#_Toc173662684)

[d. Chuyển đổi kiểu dữ liệu 19](#_Toc173662685)

[e. Xử lý lỗi định dạng 20](#_Toc173662686)

[f. Tách cột 22](#_Toc173662687)

[g. Chuẩn hóa giá trị Null 23](#_Toc173662688)

[h. API chuẩn hóa cột Country 24](#_Toc173662689)

[i. Xử lý outlier 25](#_Toc173662690)

[j. Điền dữ liệu Null cho cột Age 26](#_Toc173662691)

[k. Ràng buộc quan hệ cho cơ sở dữ liệu 26](#_Toc173662692)

[3.4.2. Lưu trữ dữ liệu Silver layer bằng Azure Blob Storage 27](#_Toc173662693)

[3.4.3. Xử lý dữ liệu đưa vào Gold layer trên Databricks 28](#_Toc173662694)

[3.4.4. Lưu trữ dữ liệu Gold layer bằng Azure Blob Storage 29](#_Toc173662695)

[3.4.5. Phân tích bằng Power BI 30](#_Toc173662696)

[3.5. Kết luận 31](#_Toc173662697)

[CHƯƠNG 4. TRIỂN KHAI THỰC NGHIỆM 32](#_Toc173662698)

[4.1. Kết luận 32](#_Toc173662699)

[4.1.1. Tóm tắt các hoạt động đã thực hiện 32](#_Toc173662700)

[4.1.2. Đánh giá kết quả đạt được 32](#_Toc173662701)

[4.1.3. Những bài học và kinh nghiệm rút ra 32](#_Toc173662702)

[4.1.4. Nhận xét về môi trường làm việc 33](#_Toc173662703)

[4.1.5. Kết luận tổng quát 33](#_Toc173662704)

[4.2. Đề xuất 33](#_Toc173662705)

[KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 34](#_Toc173662706)

[TÀI LIỆU THAM KHẢO 35](#_Toc173662707)

[CHECK LIST CỦA BÁO CÁO 36](#_Toc173662708)

[PHỤ LỤC 38](#_Toc173662709)

# DANH MỤC HÌNH ẢNH

[Hình 1.1‑1. Logo TMA Solutions Bình Định 4](#_Toc173877215)

[Hình 3.4‑1. Luồng dữ liệu của project 15](#_Toc173877216)

[Hình 3.4‑2. Đọc dữ liệu Books 16](#_Toc173877217)

[Hình 3.4‑3. Đọc dữ liệu Users 16](#_Toc173877218)

[Hình 3.4‑4. Đọc dữ liệu Ratings 17](#_Toc173877219)

[Hình 3.4‑5. Xóa cột bảng Books 17](#_Toc173877220)

[Hình 3.4‑6. Xóa dữ liệu trùng ISNB bảng Books 18](#_Toc173877221)

[Hình 3.4‑7. Xóa dữ liệu trùng User-ID bảng Users 18](#_Toc173877222)

[Hình 3.4‑8. Xóa dữ liệu trùng cặp khóa ISBN & User-ID bảng Ratings 19](#_Toc173877223)

[Hình 3.4‑9. Chuyển kiểu dữ liệu chỉ số đánh giá sách bảng Ratings 19](#_Toc173877224)

[Hình 3.4‑10. Chuyển kiểu dữ liệu Age của bảng Users 20](#_Toc173877225)

[Hình 3.4‑11. Dữ liệu cột Book-Title 20](#_Toc173877226)

[Hình 3.4‑12. Định dạng cột Book-Title, Publisher, Book-Author 21](#_Toc173877227)

[Hình 3.4‑13. Định dạng cột Location 21](#_Toc173877228)

[Hình 3.4‑14. Xử lý kí tự đặc biệt bảng Books 22](#_Toc173877229)

[Hình 3.4‑15. Xử lý kí tự đặc biệt bảng Users 22](#_Toc173877230)

[Hình 3.4‑16. Tách cột Location 23](#_Toc173877231)

[Hình 3.4‑17. Trước khi xử lý Null Country bảng Users 23](#_Toc173877232)

[Hình 3.4‑18. Xử lý Null Country bảng Users 24](#_Toc173877233)

[Hình 3.4‑19. Cột Country chưa chuẩn hóa 24](file:///D:\Downloads\Ha_DuongThiKieuDoan_47K21.2.docx#_Toc173877234)

[Hình 3.4‑20. Code chuẩn hóa cột Country 25](#_Toc173877235)

[Hình 3.4‑21. Sau khi chuẩn hóa 25](#_Toc173877236)

[Hình 3.4‑22. Xử lý outlier cột Age bảng Users 26](#_Toc173877237)

[Hình 3.4‑23. Ràng buộc dữ liệu 27](#_Toc173877238)

[Hình 3.4‑24. Liên kết và lưu vào Azure Blob Storage 27](#_Toc173877239)

[Hình 3.4‑25. Kết quả đưa lên Azure Blob Storage 28](#_Toc173877240)

[Hình 3.4‑26. Tạo thêm cột Average\_Ratings 29](#_Toc173877241)

[Hình 3.4‑27. Kết nối và lưu trữ Gold layer lên Blob Storage 29](#_Toc173877242)

[Hình 3.4‑28. Kết quả lưu trữ Gold layer 29](#_Toc173877243)

[Hình 3.4‑29. Kết quả lưu trữ Gold layer 30](#_Toc173877244)

[Hình 3.4‑30. Dashboard 31](#_Toc173877245)

# DANH MỤC BẢNG BIỂU

[Bảng 1. Giải thích bảng Books 12](#_Toc173957406)

[Bảng 2. Giải thích bảng Users 13](#_Toc173957407)

[Bảng 3. Giải thích bảng Ratings 13](#_Toc173957408)

# DANH MỤC CÁC TỪ VIẾT TẮT

**API** : Application Programming Interface

# LỜI MỞ ĐẦU

1. **Mục tiêu của đề tài**

* Xây dựng và triển khai một pipeline trên nền tảng đám mây.
* Tối ưu hóa quy trình xử lý dữ liệu.
* Tích hợp dữ liệu từ nhiều nguồn khác nhau.
* Đánh giá và giám sát pipeline.

1. **Đối tượng và phạm vi nghiên cứu**

2.1. Đối tượng

* Pipeline dữ liệu: Tập trung vào việc xây dựng, triển khai lên nền tảng đám mây và phân tích dữ liệu để hỗ trợ quá trình kinh doanh của doanh nghiệp.
* Nền tảng đám mây: Azure Data Blob, DataBricks.
* Nguồn dữ liệu CSV

2.2. Phạm vi nghiên cứu

* Thiết kế phương pháp xử lý dữ liệu: Tập trung vào các phương pháp xử lý dữ liệu.
* Lưu trữ dữ liệu: Sử dụng Azure để lưu trữ dữ liệu lên nền tảng đám mây.
* Báo cáo và phân tích: Xây dựng các báo cáo và phân tích dữ liệu cho khách hàng.

1. **Kết cấu của đề tài**

Đề tài được tổ chức gồm phần mở đầu, 4 chương nội dung và phần kết luận...

* Mở đầu
* **Chương 1**: **Tổng quan về công ty tma solutions bình định**
* **Chương 2**: **Lý thuyết về Data Engineer**
* **Chương 3: Phân tích thiết kế database**
* **Chương 4: Triển khai thực nghiệm**
* Kết luận và hướng phát triển

## TỔNG QUAN VỀ DOANH NGHIỆP

### Giới thiệu tổng quát về doanh nghiệp thực tập

#### Thông tin chung về TMA Solutions Bình Định.

Được thành lập vào năm 1997, TMA trở thành tập đoàn công nghệ hàng đầu tại Việt Nam, phát triển vững mạnh cùng đội ngũ gần 4,000 kỹ sư tài năng. TMA tự hào khi hợp tác với các khách hàng, tập đoàn công nghệ hàng đầu từ hơn 30 quốc gia khác nhau trên toàn thế giới.

Hiện nay, TMA trải rộng thị trường với 7 chi nhánh tại Việt Nam, trong đó 6 chi nhánh tại thành phố Hồ Chí Minh và 1 chi nhánh tại thành phố Quy Nhơn, Bình Định. Ngoài ra, TMA còn có mạng lưới quốc tế với 6 chi nhánh tại các quốc gia khác nhau bao gồm Mỹ, Úc, Canada, Đức, Nhật Bản và Singapore.

Tại thời điểm tháng 6 năm 2018, TMA đã có bước chuyển mình khi mở văn phòng đầu tiên tại TP Quy Nhơn, Bình Định. Hiểu được tâm tư, nguyện vọng của các kỹ sư trẻ đang làm việc tại TP HCM mong muốn được trở về quê hương sinh sống, sau 5 năm, TMA Bình Định vui mừng là nơi hơn 600 nhân viên lựa chọn cùng công ty đóng góp xây dựng quê hương.

Với sự phát triển nhanh chóng, tháng 8 năm 2018, TMA Bình Định khởi công xây dựng Công viên sáng tạo phần mềm TMA (TMA Innovation Park - TIP) với diện tích hơn 10 hecta tọa lạc tại Thung lũng Sáng tạo Quy Nhơn (Quy Nhon Innovation Park – QNIVY) với vốn đầu tư hàng trăm tỷ đồng.

Vinh dự là trung tâm sáng tạo phần mềm đầu tiên tại Thung lũng Sáng tạo tại Quy Nhơn, Bình Định, Công viên Sáng tạo TMA mang trong mình sứ mệnh trở thành trung tâm phát triển phần mềm và công nghệ cao hàng đầu tại miền Trung, góp phần quan trọng đưa Thung lũng Sáng tạo Quy Nhơn thành một trong những điểm sáng khi nhắc đến nền công nghệ 5.0 tại Việt Nam.



Hình 1.1‑1. Logo TMA Solutions Bình Định

#### Lĩnh vực hoạt động

* Ngân hàng và tài chính
* Viễn thông
* Chính phủ và dịch vụ công
* Sản xuất và kinh doanh
* Giáo dục và đào tạo

### Tổng quan về vị trí việc làm

#### Định nghĩa về Data Engineer [1]

Data Engineer là một chuyên gia công nghệ thông tin chuyên thiết kế, xây dựng và quản lý các hệ thống xử lý dữ liệu. Vai trò của họ là tối ưu hóa việc thu thập, lưu trữ và phân tích dữ liệu, nhằm hỗ trợ các tổ chức trong việc ra quyết định và vận hành.

#### Nhiệm vụ của một Data Engineer

Nhiệm vụ của một Data Engineer bao gồm nhiều công việc cụ thể nhằm thiết kế, xây dựng và duy trì hệ thống dữ liệu. Dưới đây là các nhiệm vụ chính của một Data Engineer:

- Thiết Kế Kiến Trúc Dữ Liệu

- Phát Triển và Duy Trì Các Pipeline Dữ Liệu

- Tối Ưu Hóa Hệ Thống Dữ Liệu

- Quản Lý và Bảo Mật Dữ Liệu

- Hỗ Trợ Phân Tích Dữ Liệu

- Giám Sát và Bảo Trì Hệ Thống

- Phát Triển và Thử Nghiệm Công Cụ Dữ Liệu

- Hợp Tác với Các Bộ Phận Khác

- Cập Nhật và Nâng Cao Kiến Thức

### Các kỹ năng cần có để trở thành một Data Engineer

#### Ngôn ngữ lập trình

* Python: Một trong những ngôn ngữ phổ biến nhất cho công việc về dữ liệu.
* Java hoặc Scala: Thường được sử dụng trong các hệ thống big data.
* SQL: Cần thiết để truy vấn và thao tác với cơ sở dữ liệu.

#### Cơ sở dữ liệu

* SQL Databases: Như SQL, MySQL, PostgreSQL, hay Oracle.
* NoSQL Databases: Như MongoDB, Cassandra, hay HBase.

#### Big Data Technologies

* Hadoop: Framework mã nguồn mở để xử lý và lưu trữ lượng lớn dữ liệu.
* Spark: Công cụ xử lý dữ liệu nhanh chóng, thường được sử dụng thay thế hoặc cùng với Hadoop.
* Kafka: Hệ thống nhắn tin phân tán được sử dụng để xây dựng các ứng dụng dữ liệu streaming.

#### ETL (Extract, Transform, Load)

* Công cụ ETL: Như Apache NiFi, Talend, và Informatica.
* Quy trình ETL: Hiểu biết về cách trích xuất, biến đổi, và tải dữ liệu.

#### Cloud Platforms

* AWS: Các dịch vụ như S3, Redshift, EMR.
* Google Cloud: BigQuery, Dataflow, Pub/Sub.
* Microsoft Azure: Azure Data Lake, Azure SQL Data Warehouse.

#### Data Warehousing

* Hiểu biết về kiến trúc và thiết kế kho dữ liệu, như Snowflake, Redshift, hoặc Google BigQuery.

### Cơ hội nghề nghiệp của Data Engineer

* ETL Developer: Chuyên về phát triển và duy trì quy trình ETL (Extract, Transform, Load) để chuyển đổi dữ liệu từ nguồn vào hệ thống lưu trữ.
* Data Architect: Thiết kế và xây dựng kiến trúc dữ liệu, đảm bảo tính mở rộng và hiệu quả.
* Data Pipeline Engineer: Xây dựng và duy trì các pipeline dữ liệu để chuyển dữ liệu từ nguồn đến hệ thống đích một cách hiệu quả.
* Database Administrator (DBA): Quản lý và tối ưu hóa cơ sở dữ liệu, đảm bảo dữ liệu luôn sẵn sàng và bảo mật.
* Big Data Engineer: Chuyên về các công nghệ big data như Hadoop, Spark, và Kafka để xử lý và phân tích dữ liệu lớn.

### Kết luận

Trong chương này, chúng ta đã cung cấp một cái nhìn tổng quan về doanh nghiệp thực tập, bao gồm thông tin chung về TMA Solutions Bình Định, các lĩnh vực hoạt động, và định nghĩa về vị trí Data Engineer. Chúng ta cũng đã trình bày các nhiệm vụ, kỹ năng cần có, và cơ hội nghề nghiệp của một Data Engineer. Qua đó, người đọc có thể hiểu rõ hơn về vai trò và triển vọng của công việc này trong ngành công nghệ thông tin.

## CƠ SỞ LÝ THUYẾT VỀ DATA ENGINEER

### Giới thiệu về Data Engineering [2]

**Định nghĩa**: Data Engineering là quá trình thiết kế, xây dựng và quản lý hệ thống để thu thập, lưu trữ, và phân tích dữ liệu. Data Engineer chịu trách nhiệm tạo ra cơ sở hạ tầng dữ liệu, bao gồm cả việc xây dựng các đường ống dữ liệu và đảm bảo rằng dữ liệu được chuẩn bị và sẵn sàng cho việc phân tích.

**Tầm quan trọng**: Với sự gia tăng của dữ liệu lớn và chuyển đổi số, vai trò của Data Engineer trở nên vô cùng quan trọng trong việc giúp các tổ chức xử lý và khai thác dữ liệu một cách hiệu quả. Data Engineer đảm bảo rằng dữ liệu luôn sẵn sàng, chính xác và có thể truy cập được cho các nhà phân tích và khoa học dữ liệu.

### Các khái niệm cơ bản [2]

**Kho dữ liệu (Data Warehouse):** Kho dữ liệu là một hệ thống lưu trữ dữ liệu được tối ưu hóa cho việc truy vấn và phân tích dữ liệu. Dữ liệu trong kho dữ liệu thường được tổng hợp từ nhiều nguồn khác nhau và được tổ chức theo cách giúp dễ dàng truy vấn và phân tích.

**Hệ thống ETL (Extract, Transform, Load):** ETL là quá trình trích xuất (Extract) dữ liệu từ các nguồn khác nhau, biến đổi (Transform) dữ liệu thành định dạng phù hợp, và tải (Load) dữ liệu vào hệ thống lưu trữ như kho dữ liệu. ETL là quy trình quan trọng trong việc chuẩn bị dữ liệu cho phân tích.

**Điện toán đám mây (Cloud Computing):** Là mô hình cung cấp dịch vụ điện toán qua internet, cho phép người dùng truy cập vào tài nguyên như lưu trữ dữ liệu, máy chủ, và phần mềm từ xa, mà không cần đầu tư vào phần cứng hoặc phần mềm cục bộ.

**DataFrame:** Là một cấu trúc dữ liệu được sử dụng để lưu trữ và xử lý các bộ dữ liệu lớn. Nó là một cấu trúc dữ liệu dạng hai chiều, giống như một bảng trong cơ sở dữ liệu quan hệ hoặc một bảng tính Excel, bao gồm các hàng và các cột.

**Raw Layer:** Là giai đoạn đầu tiên, nơi dữ liệu được lưu trữ dưới dạng gốc, chưa qua xử lý, bao gồm cả giá trị lỗi và không nhất quán. Đây là dữ liệu nguyên bản thu thập từ nhiều nguồn khác nhau.

**Silver Layer:** Chứa dữ liệu đã được xử lý sơ bộ, làm sạch và biến đổi cơ bản. Các bước xử lý bao gồm loại bỏ dữ liệu trùng lặp, sửa lỗi, xử lý giá trị thiếu và chuẩn hóa dữ liệu. Lớp này có chất lượng tốt hơn và có thể được sử dụng cho các phân tích sơ bộ.

**Gold layer:** Là giai đoạn cuối cùng, nơi dữ liệu đã được xử lý hoàn toàn và sẵn sàng cho các phân tích cao cấp, báo cáo hoặc học máy. Dữ liệu trong lớp này đã được tối ưu hóa, tích hợp và chuẩn hóa để đảm bảo tính toàn vẹn và nhất quán cao nhất.

### Lý thuyết liên quan

**Cơ sở dữ liệu quan hệ (SQL)**: Structured Query Language là ngôn ngữ tiêu chuẩn để quản lý và truy vấn các cơ sở dữ liệu quan hệ. SQL cho phép thực hiện các thao tác như thêm, xóa, sửa và truy vấn dữ liệu.

**OLTP (Online Transaction Processing):** Là một loại hệ thống xử lý giao dịch trực tuyến, được thiết kế để quản lý và xử lý một số lượng lớn các giao dịch cơ sở dữ liệu ngắn và đơn giản mà người dùng thực hiện.

**OLAP (Online Analytical Processing**): Là một công nghệ xử lý dữ liệu được thiết kế để hỗ trợ các hoạt động phân tích và báo cáo dữ liệu phức tạp, cung cấp khả năng truy vấn và phân tích dữ liệu nhanh chóng từ nhiều góc độ khác nhau. OLAP thường được sử dụng trong các công cụ phân tích dữ liệu và báo cáo để hỗ trợ việc ra quyết định chiến lược trong các doanh nghiệp.

### Công nghệ và công cụ

**Spark**: là một công cụ xử lý dữ liệu phân tán nhanh chóng, cung cấp các API cho Java, Scala, Python, và R. Spark hỗ trợ xử lý dữ liệu thời gian thực và cung cấp thư viện cho machine learning và xử lý đồ thị.

**PySpark:** là một giao diện cho Apache Spark trong Python, cho phép sử dụng các API của Spark với Python, giúp các nhà khoa học dữ liệu và kỹ sư dữ liệu có thể xử lý và phân tích dữ liệu lớn một cách hiệu quả.

**Python**: là ngôn ngữ lập trình thông dịch, phổ biến trong Data Engineering nhờ vào tính đơn giản và mạnh mẽ, cùng với nhiều thư viện hỗ trợ như Pandas, NumPy, và PySpark.

### Kết luận

Chương này giới thiệu về Data Engineering và các khái niệm cơ bản liên quan. Chúng ta đã thảo luận về lý thuyết liên quan cũng như công nghệ và công cụ cần thiết cho Data Engineering. Những thông tin này cung cấp nền tảng lý thuyết quan trọng để hiểu rõ hơn về vai trò và trách nhiệm của một Data Engineer.

## PHÂN TÍCH THIẾT KẾ DATABASE

### Khảo sát yêu cầu và phân tích

Trong giai đoạn khảo sát yêu cầu và phân tích cho đề tài tạo đường ống xử lý dữ liệu để đưa ra chiến lược kinh doanh cho một nhà xuất bản sách, tôi đã tham gia các cuộc họp giúp xác định rõ các nguồn dữ liệu cần thiết. Chúng tôi cũng làm rõ các loại dữ liệu cần trích xuất bao gồm thông tin khách hàng, doanh số bán hàng theo thể loại sách, và hiệu quả của các chiến dịch marketing. Dựa trên các yêu cầu này, tôi đã biên soạn tài liệu chi tiết mô tả mục tiêu của hệ thống ETL, các bước trích xuất, biến đổi và tải dữ liệu, cùng với các chỉ số phân tích cần thiết như cuốn sách được đánh giá cao, loại sách ưa thích theo từng độ tuổi, và xu hướng mua sách của khách hàng. Tài liệu này được xem xét, phản hồi và phê duyệt bởi tất cả các bên liên quan trước khi tiến hành thiết kế và triển khai hệ thống.

### Tổng quan về dữ liệu

#### Giới thiệu tổng quan về dữ liệu

Cơ sở dữ liệu bao gồm ba file CSV: Books, Ratings, và Users. File Books chứa thông tin chi tiết về các cuốn sách như tiêu đề, tác giả, năm xuất bản và nhà xuất bản. File Ratings ghi nhận các đánh giá của người dùng về sách, bao gồm điểm số mà họ chấm cho từng cuốn sách. File Users chứa thông tin về người dùng, bao gồm vị trí địa lý và độ tuổi. Các file này được liên kết với nhau qua mã ISBN của sách và ID người dùng, cho phép phân tích hành vi đọc sách và đánh giá sách của người dùng, từ đó hỗ trợ nhà xuất bản đưa ra các chiến lược kinh doanh dựa trên dữ liệu thực tế.

#### Các bảng của dữ liệu

##### Bảng Books

|  |  |  |
| --- | --- | --- |
| ***Cột*** | ***Loại dữ liệu*** | ***Giải thích*** |
| ISBN | String | Mã số tiêu chuẩn quốc tế, là một định danh duy nhất cho mỗi cuốn sách. |
| Book-Title | String | Tên của cuốn sách. |
| Book-Author | String | Tác giả của cuốn sách. |
| Year-Of-Publication | Date (Year) | Năm xuất bản của cuốn sách. |
| Publisher | String | Nhà xuất bản của cuốn sách. |
| Image-URL-S | String | URL của hình ảnh bìa sách kích thước nhỏ. |
| Image-URL-M | String | URL của hình ảnh bìa sách kích thước trung bình. |
| Image-URL-L | String | URL của hình ảnh bìa sách kích thước lớn. |

Bảng 1. Giải thích bảng Books

##### Bảng Ratings

|  |  |  |
| --- | --- | --- |
| ***Cột*** | ***Kiểu dữ liệu*** | ***Giải thích*** |
| User-ID | String | Định danh duy nhất của người dùng. |
| ISBN | String | Mã số tiêu chuẩn quốc tế cho sách, liên kết với file Books. |
| Book-Rating | String | Điểm đánh giá của người dùng dành cho cuốn sách, có thể nằm trong khoảng từ 0 đến 10. |

Bảng 2. Giải thích bảng Users

##### Bảng Users

|  |  |  |
| --- | --- | --- |
| ***Cột*** | ***Kiểu dữ liệu*** | ***Giải thích*** |
| User-ID | String | Định danh duy nhất của người dùng. |
| Location | String | Vị trí địa lý của người dùng, thường bao gồm thành phố và quốc gia. |
| Age | Int | Vị trí địa lý của người dùng, thường bao gồm thành phố và quốc gia. |

Bảng 3. Giải thích bảng Ratings

#### Mối quan hệ của các bảng trong cơ sở dữ liệu

**Books và Ratings:** Mối quan hệ giữa hai file này được thể hiện qua trường "ISBN". Mỗi cuốn sách trong file Books có thể có nhiều đánh giá trong file Ratings.

**Users và Ratings:** Mối quan hệ giữa hai file này được thể hiện qua trường "User-ID". Mỗi người dùng trong file Users có thể đánh giá nhiều cuốn sách trong file Ratings.

### Các công cụ cần dùng

#### DataBricks (Pyspark)

Databricks là một nền tảng phân tích dữ liệu dựa trên đám mây, được xây dựng trên Apache Spark. Nó cung cấp một môi trường hợp nhất để xử lý và phân tích dữ liệu lớn, kết hợp khả năng xử lý dữ liệu quy mô lớn của Spark với các tính năng quản lý, tích hợp, và hỗ trợ cộng tác nâng cao. Databricks hỗ trợ nhiều ngôn ngữ lập trình như Python, Scala, R, SQL, Java và được tích hợp sâu với các dịch vụ đám mây như AWS, Azure, và Google Cloud, cho phép người dùng tận dụng các dịch vụ lưu trữ, cơ sở dữ liệu, và các dịch vụ khác.

Trong dự án, DataBricks có vai trò quan trọng trong việc xử lý dữ liệu sơ bộ, chuyên sâu và kết nối với Azure Blob Storage để lưu trữ cơ sở dữ liệu lên môi trường đám mây.

#### Azure Blob Storage

Azure Blob Storage là một dịch vụ lưu trữ đối tượng trên nền tảng đám mây của Microsoft Azure. Nó được thiết kế để lưu trữ lượng lớn dữ liệu không có cấu trúc, bao gồm văn bản hoặc dữ liệu nhị phân.

Azure Blob Storage cũng có vai trò không kém quan trọng trong dự án như:

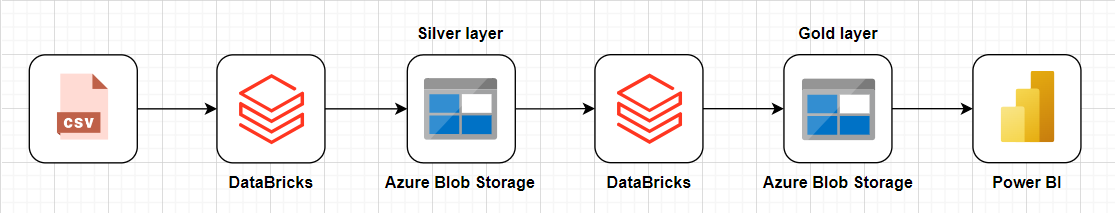
* Lưu trữ dữ liệu trung gian (Silver Layer).
* Lưu trữ dữ liệu cuối cùng (Gold Layer).
* Cung cấp dữ liệu cho công cụ trực quan hóa (Power BI).

#### Power BI

Power BI là một bộ công cụ phân tích kinh doanh được phát triển bởi Microsoft, giúp người dùng trực quan hóa dữ liệu và chia sẻ thông tin chi tiết trong toàn tổ chức hoặc nhúng chúng vào ứng dụng hoặc trang web của họ. Power BI cung cấp khả năng kết nối với hàng trăm nguồn dữ liệu, đơn giản hóa việc chuẩn bị dữ liệu và tạo ra các báo cáo và bảng điều khiển trực quan.

Đóng vai trò rất quan trọng để có thể hỗ trợ quá trình ra quyết định bằng các chức năng như tạo ra các biểu đồ, đồ thị, bảng điều khiển để minh họa dữ liệu một cách trực quan bên cạnh đó tạo ra các báo cáo tương tác và các bảng điều khiển tùy chỉnh dựa trên dữ liệu từ lớp Gold.

### Workflow của dự án



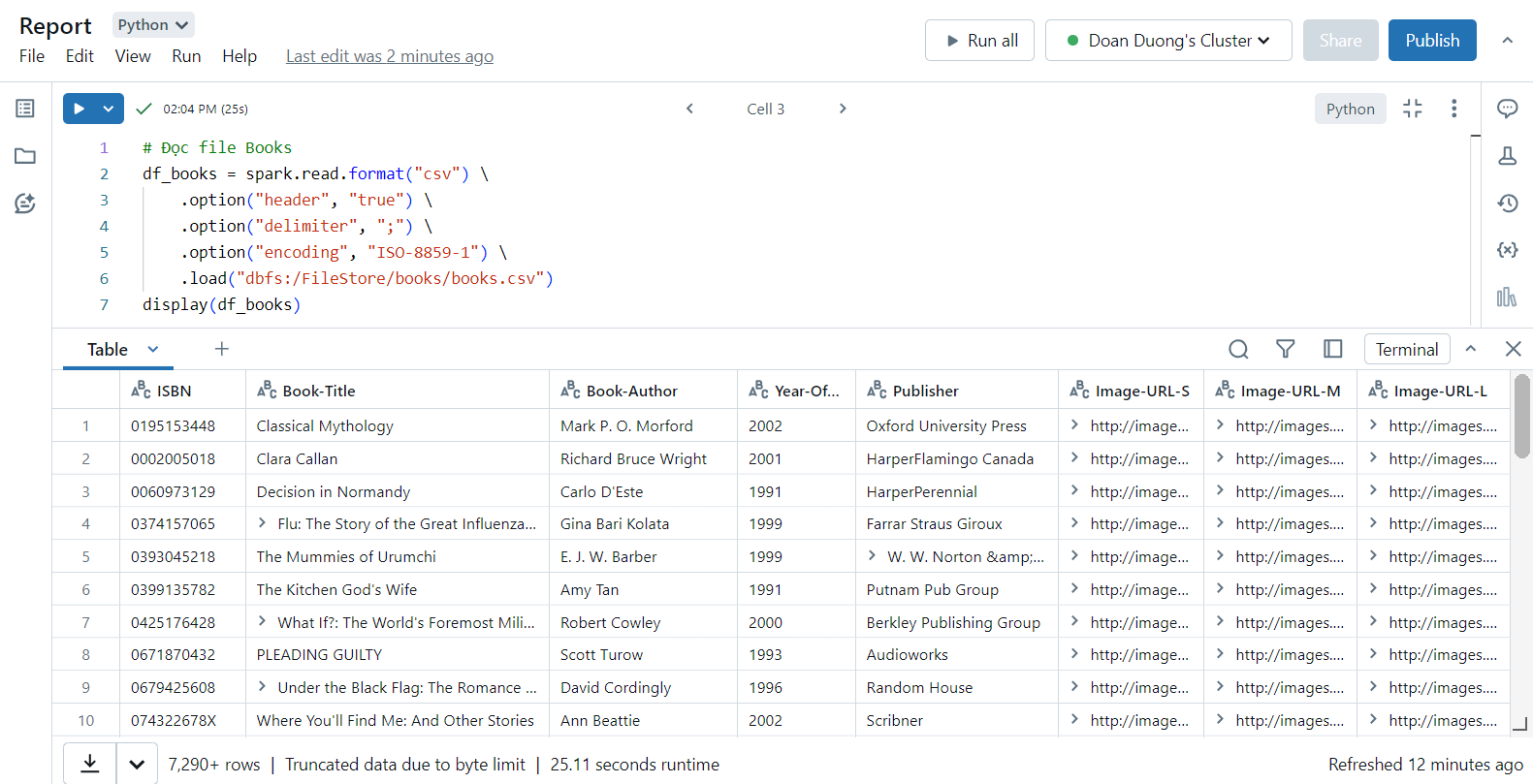
Hình 3.4‑1. Luồng dữ liệu của project

#### Xử lý dữ liệu thô bằng DataBricks

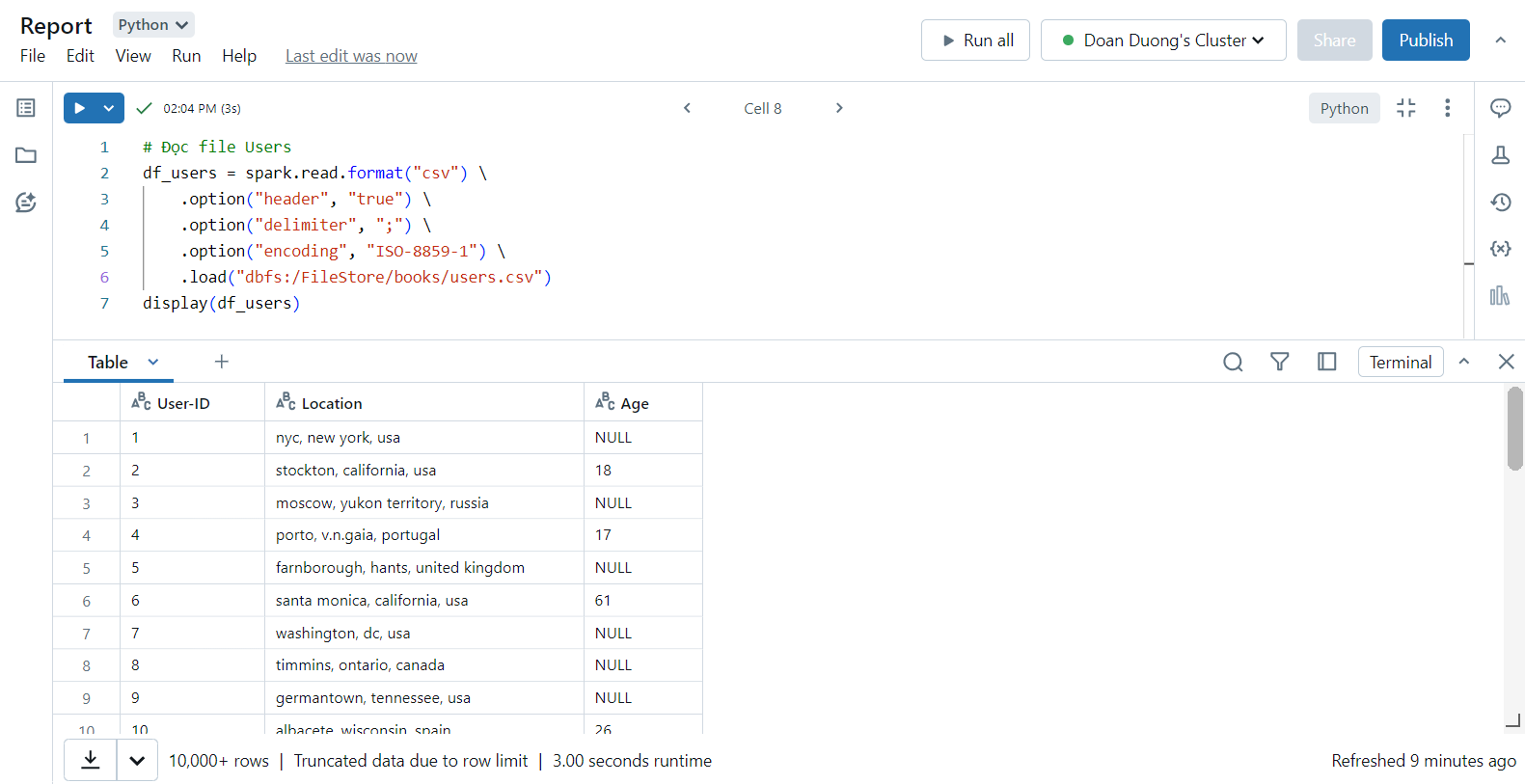
Vì dữ liệu đầu vào là ba tệp CSV với mỗi file là một bảng dữ liệu khác nhau và liên kết với nhau bằng các khóa chính, khóa ngoại nên chúng ta sẽ đọc lần lượt ba tệp dữ liệu vào DataBricks và xử lý các dữ liệu bẩn.

##### Đọc dữ liệu vào dataframe

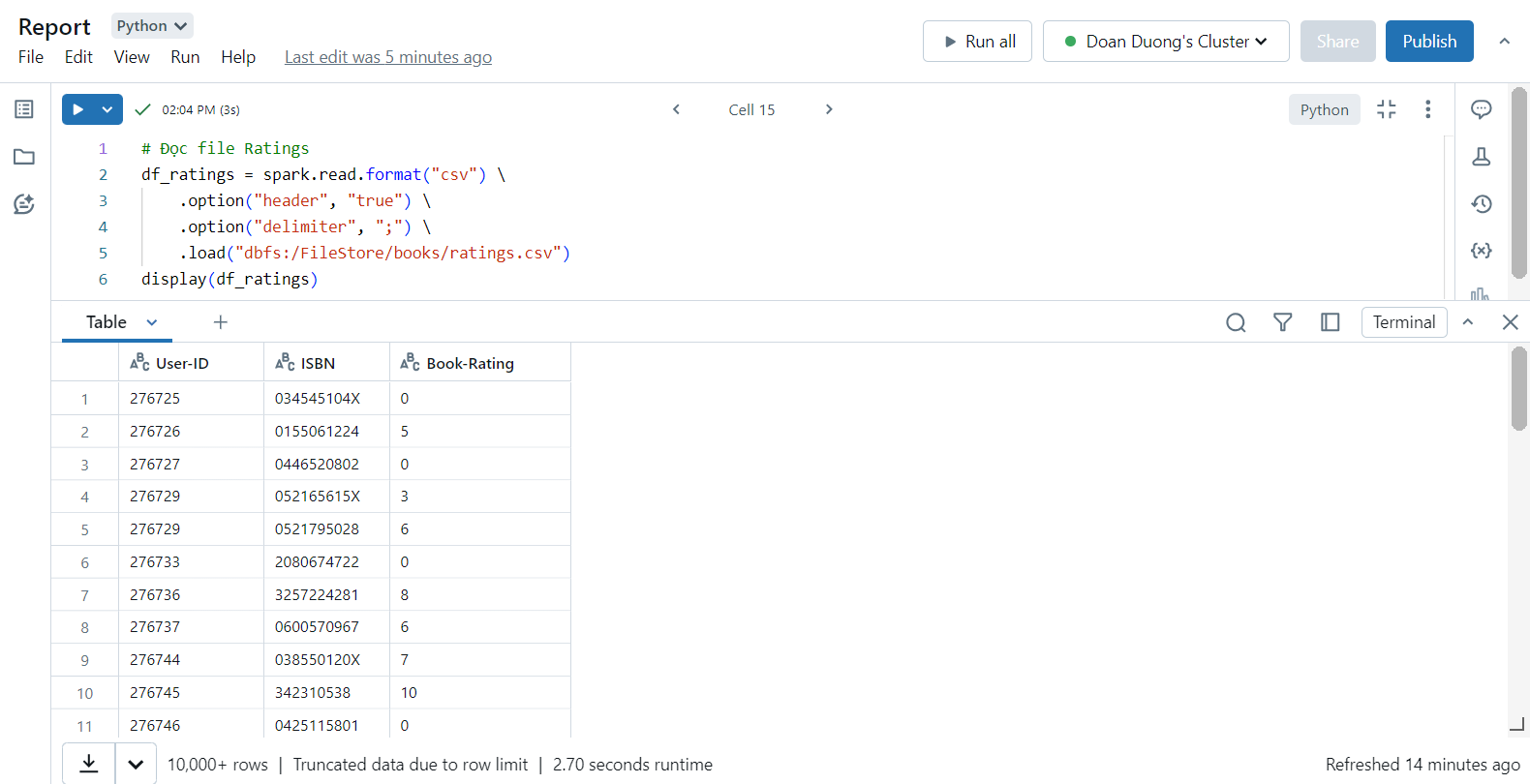
Đọc dữ liệu file Books vào dataframe df\_books để xử lý. Ở đây ta đọc bằng thư viện spark với format file là CSV. Có các tùy chọn như đọc vào có tiêu đề mục, dấu phân cách là “;”, mã hóa là loại ISO-8859-1. Tương tự với các bảng Users và Ratings.



Hình 3.4‑2. Đọc dữ liệu Books



Hình 3.4‑3. Đọc dữ liệu Users



Hình 3.4‑4. Đọc dữ liệu Ratings

##### Xóa cột dữ liệu không sử dụng để phân tích

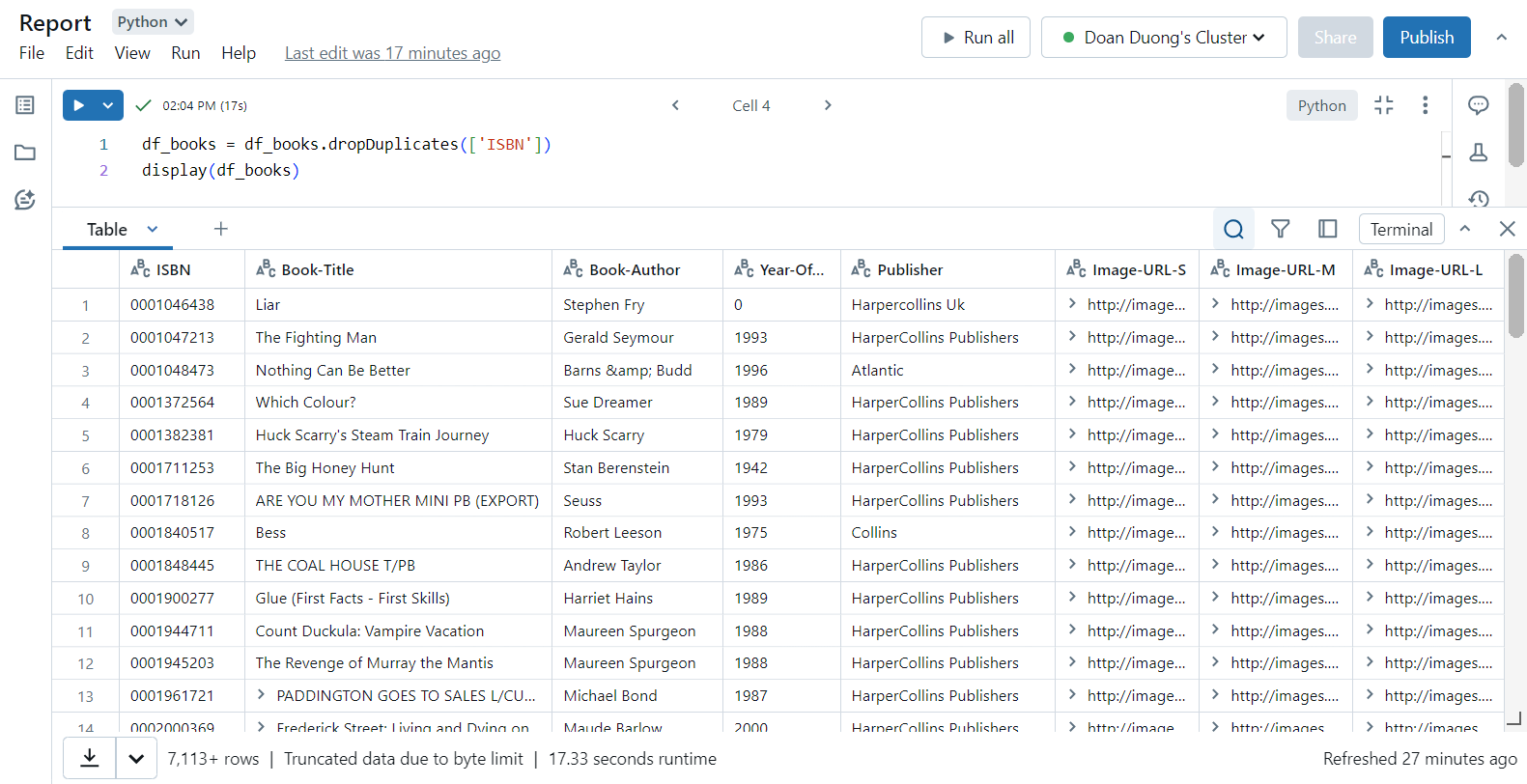
Những cột dữ liệu không phân tích sẽ không có ích cho dữ liệu và tốn nhiều bộ nhớ lưu trữ cũng như thời gian xử lý. Do đó chúng ta nên xử lý cột trước khi xử lý việc khác.



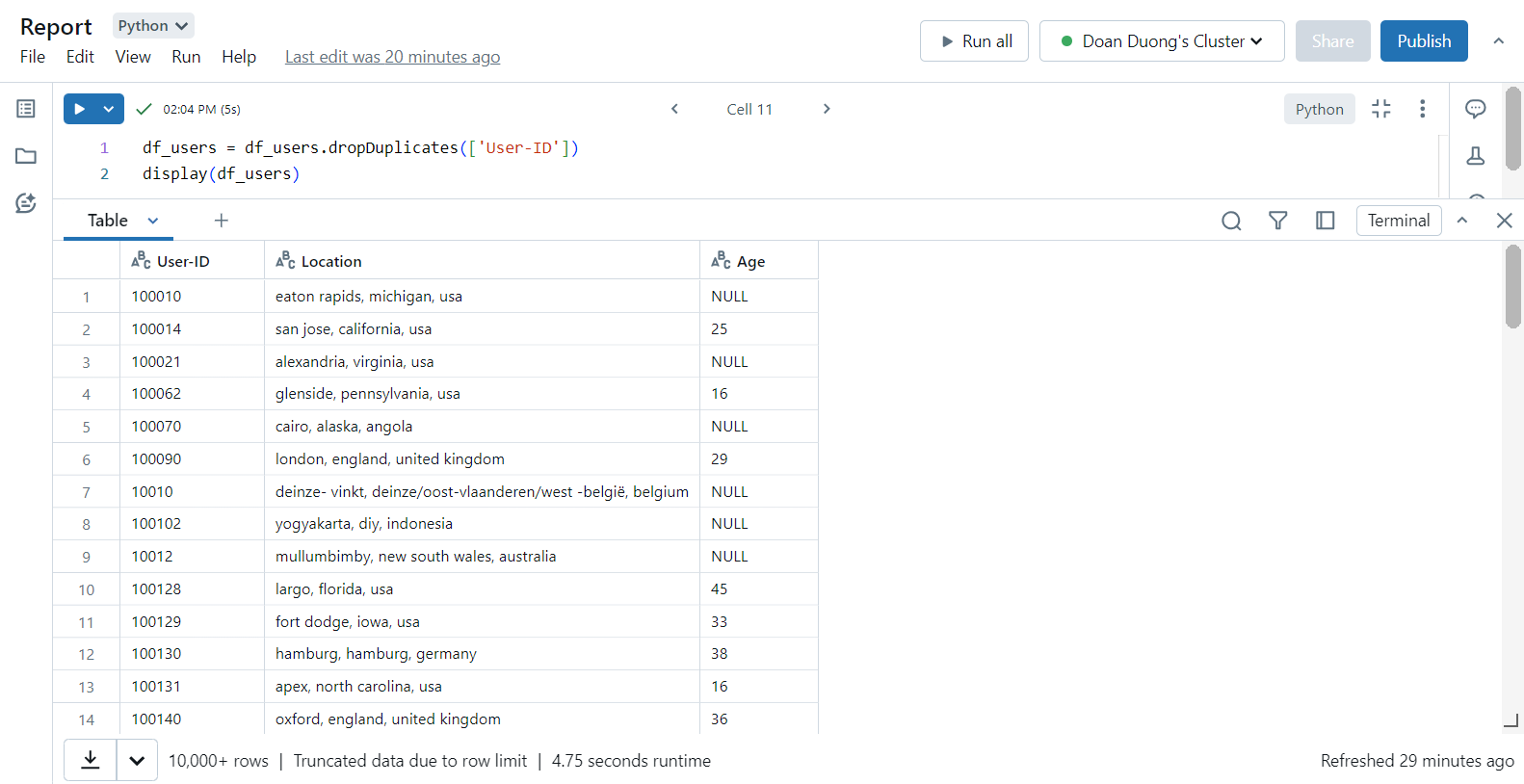
Hình 3.4‑5. Xóa cột bảng Books

##### Xóa dữ liệu trùng từ các cột khóa chính

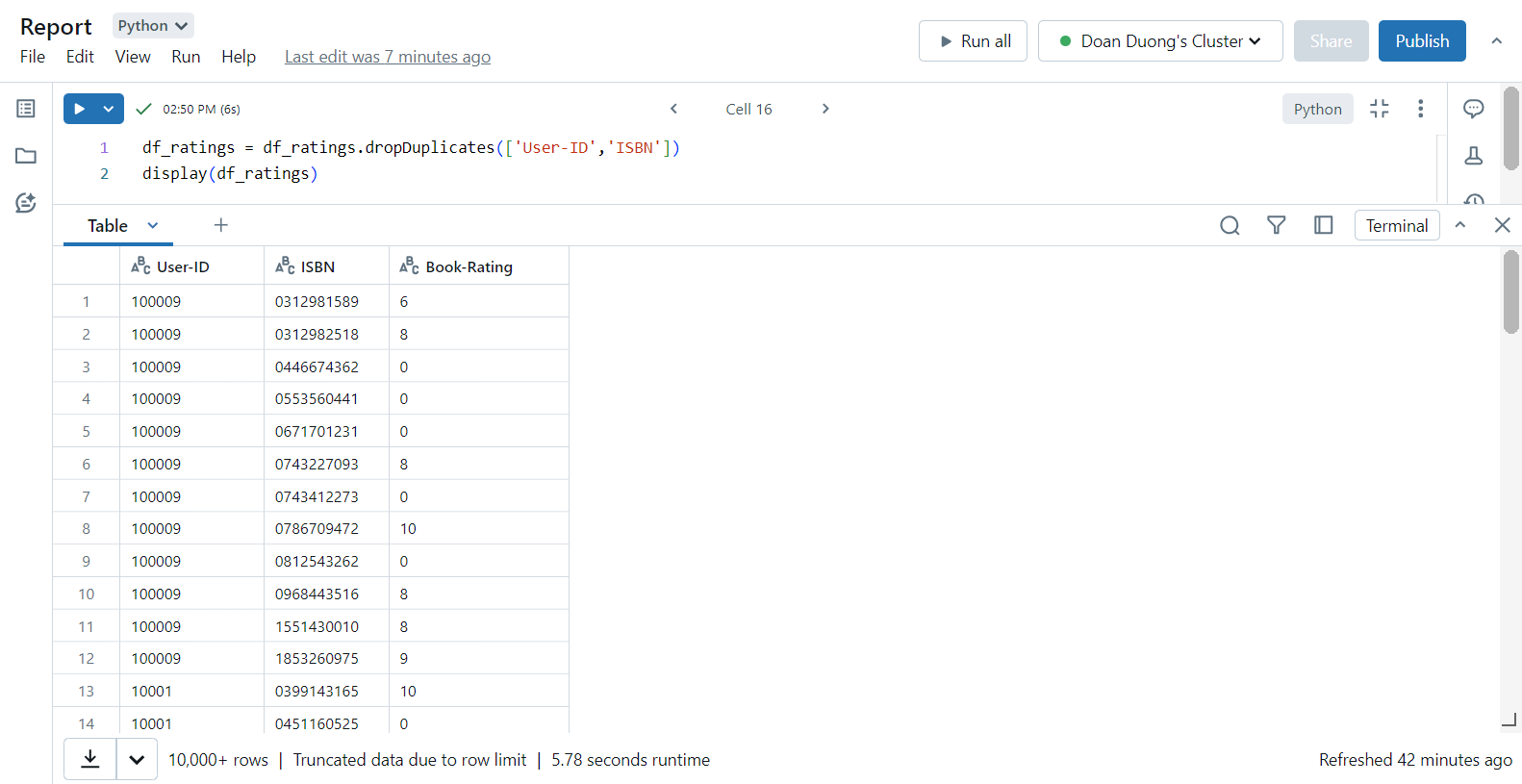
Vì trong cơ sở dữ liệu không thể trùng khóa chính nên khi xử lý chúng ta phải bỏ các bản ghi trùng khóa chính của các bảng Books và Users. Còn bảng Ratings thì vì một sách có thể có nhiều đánh giá và một người cũng có thể đánh giá nhiều sách nên khóa chính của bảng này là cặp khóa chính của hai bảng trên.



Hình 3.4‑6. Xóa dữ liệu trùng ISNB bảng Books



Hình 3.4‑7. Xóa dữ liệu trùng User-ID bảng Users



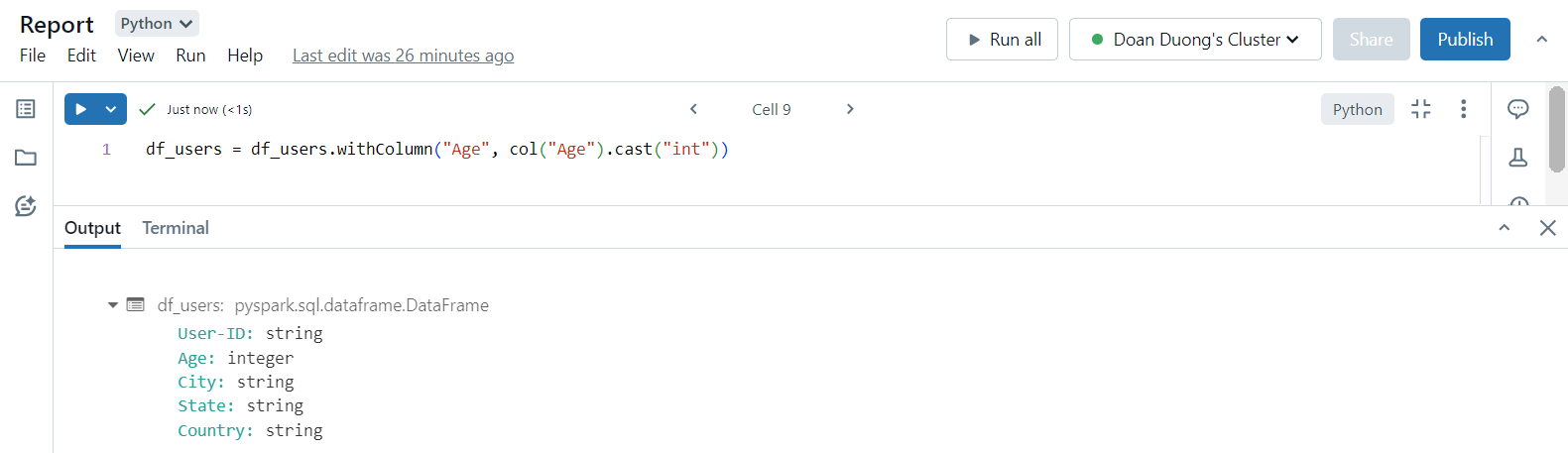
Hình 3.4‑8. Xóa dữ liệu trùng cặp khóa ISBN & User-ID bảng Ratings

##### Chuyển đổi kiểu dữ liệu

Sau khi đọc các file CSV này vào dataframe thì kiểu dữ liệu của chúng mặc định là string nên phải đổi kiểu dữ liệu cho phù hợp để phân tích các bước sau này. Trong này có cột đánh giá sách “Book – Rating” trong bảng Ratings ta sẽ đổi kiểu dữ liệu thành kiểu dữ liệu số cho dễ dàng tính toán và phân tích.



Hình 3.4‑9. Chuyển kiểu dữ liệu chỉ số đánh giá sách bảng Ratings

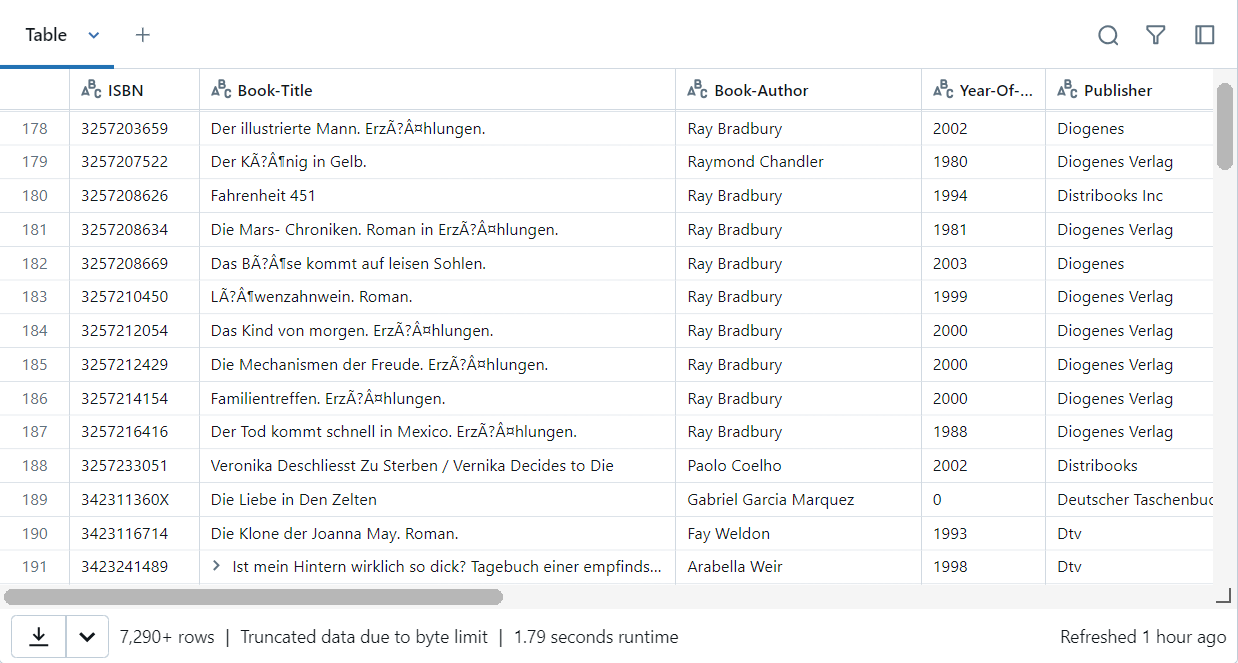


Hình 3.4‑10. Chuyển kiểu dữ liệu Age của bảng Users

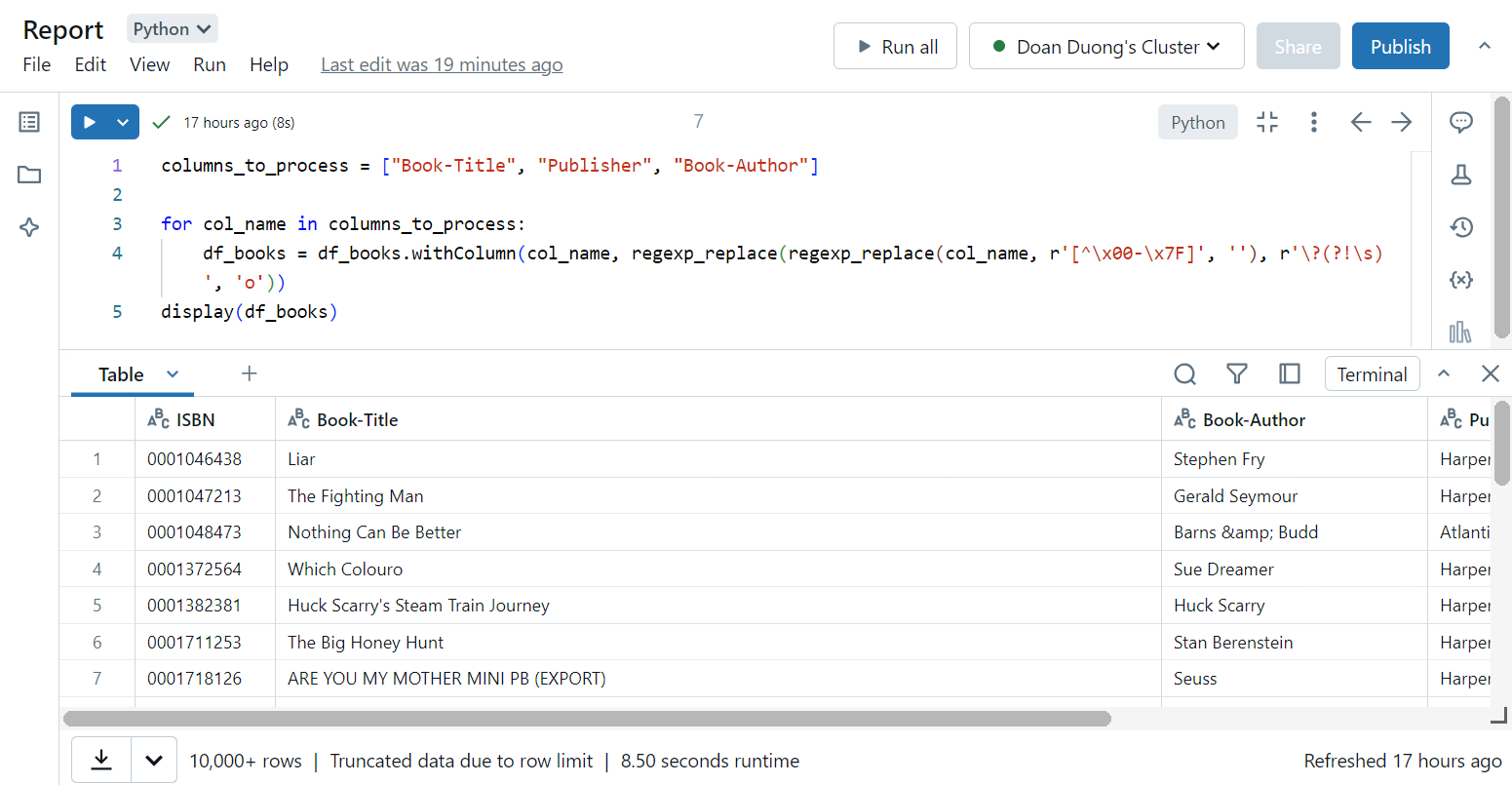
##### Xử lý lỗi định dạng

Khi đọc file vào DataBricks thì có các cột bị lỗi định dạng như cột “Book-Title”. Tuy lúc đọc có tùy chọn đọc mã hóa nhưng vẫn còn có các bản ghi lỗi định dạng. Do đó cần xử lý các lỗi định dạng đó để có một cơ sở dữ liệu được chuẩn hóa.

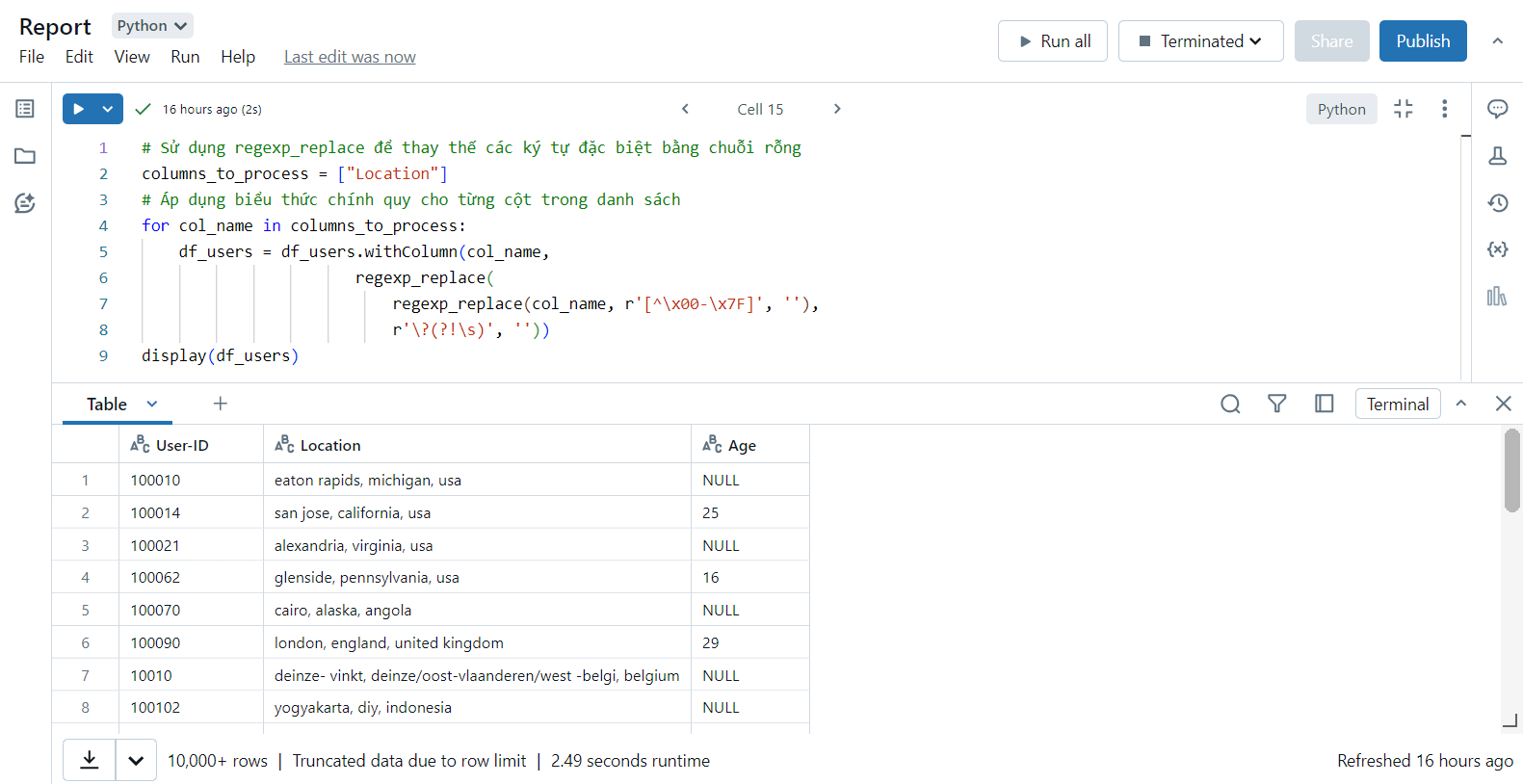
Ở đây có cột Book-Title, Publisher trong bảng Books và cột Location của bảng Users. Chúng ta sẽ loại trừ các kí tự không phải lỗi ASCII bằng kí tự rỗng.



Hình 3.4‑11. Dữ liệu cột Book-Title

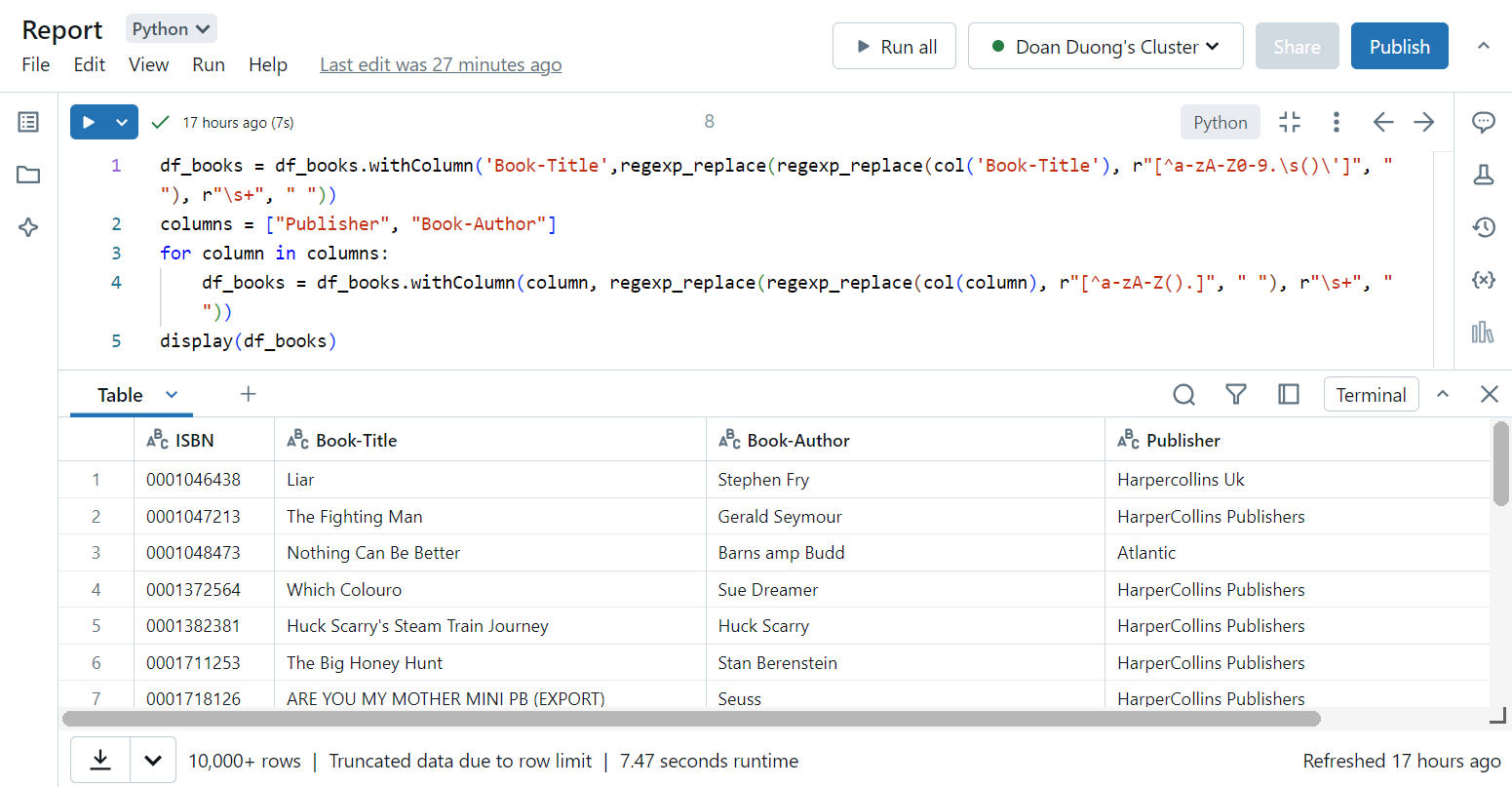


Hình 3.4‑12. Định dạng cột Book-Title, Publisher, Book-Author

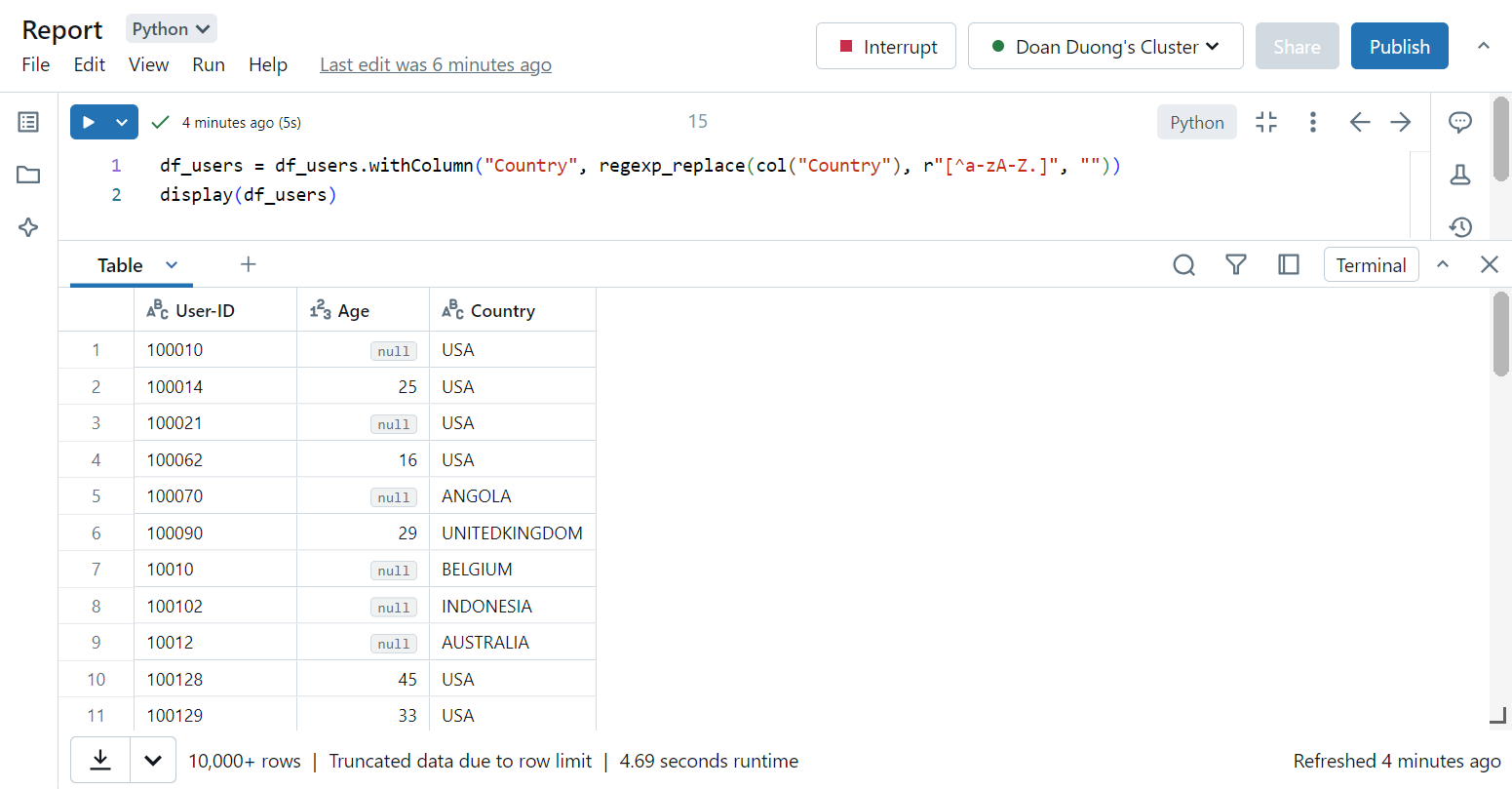


Hình 3.4‑13. Định dạng cột Location

Sau khi loại trừ các kí tự không thuộc mã ASCII thì sẽ thay thế các kí tự đặc biệt theo định nghĩa từng trường bằng khoảng trắng và thay thế các kí tự khoảng trắng liên tiếp thành một khoảng trắng để có thể chuẩn hóa thành một chuỗi hoàn chỉnh.



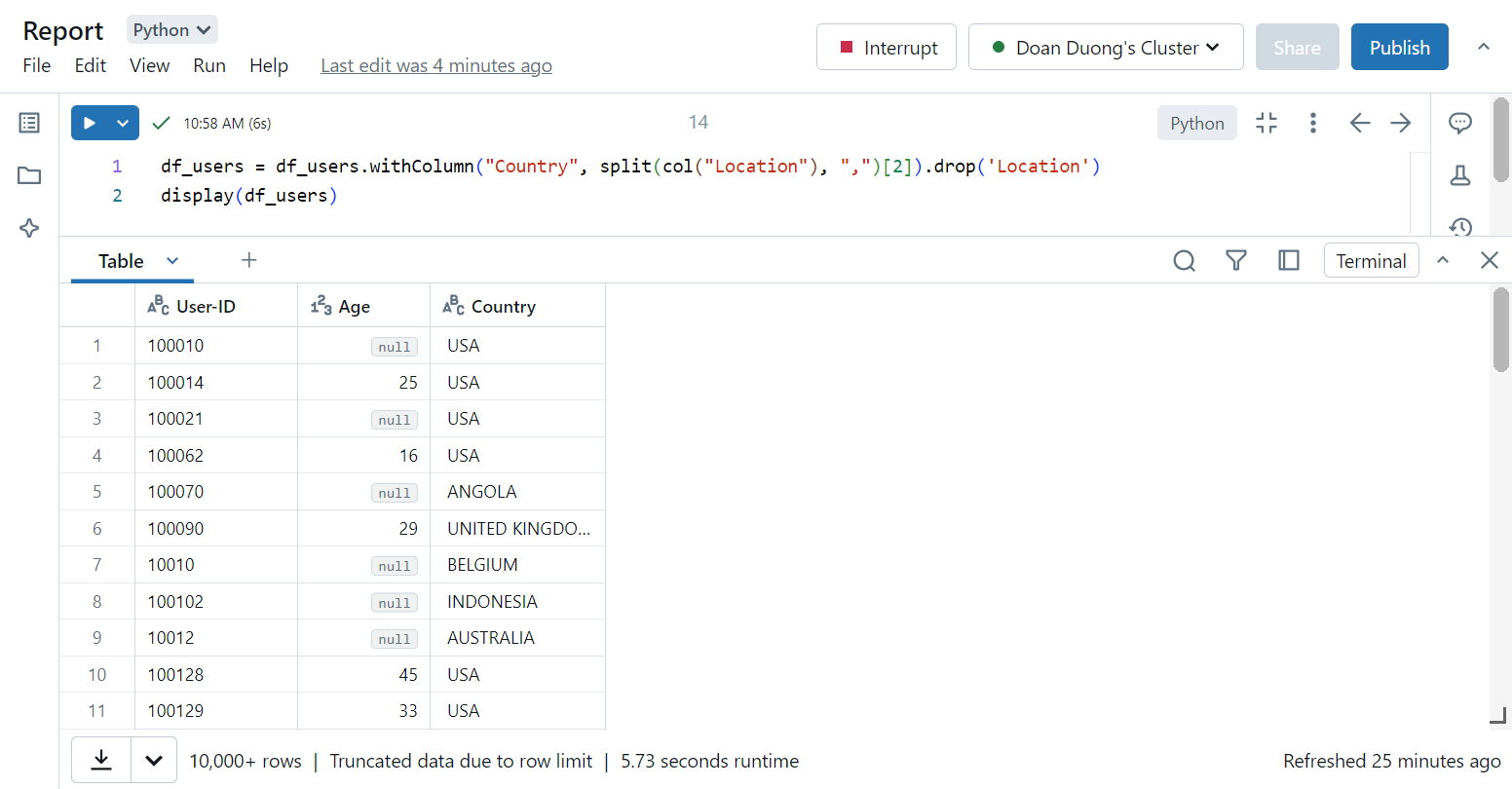
Hình 3.4‑14. Xử lý kí tự đặc biệt bảng Books



Hình 3.4‑15. Xử lý kí tự đặc biệt bảng Users

##### Tách cột

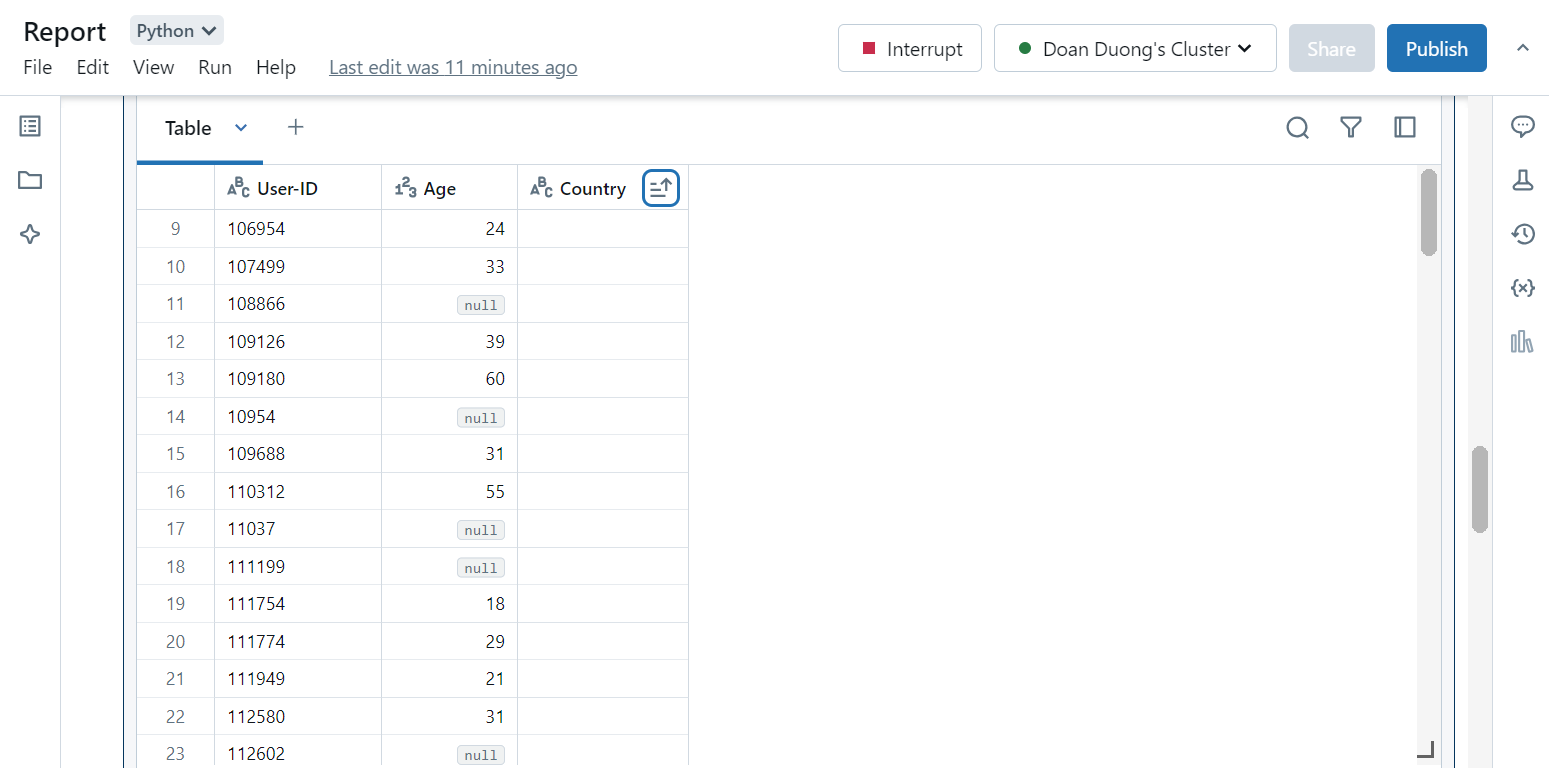
Vì cột Location có quá nhiều thông tin chưa được chuẩn hóa và mục tiêu bài phân tích chỉ cần phân tích về quốc gia nên bước này sẽ giúp lấy được thông tin Country từ Location để có thể chuẩn hóa và dễ dàng phân tích ở phần sau.



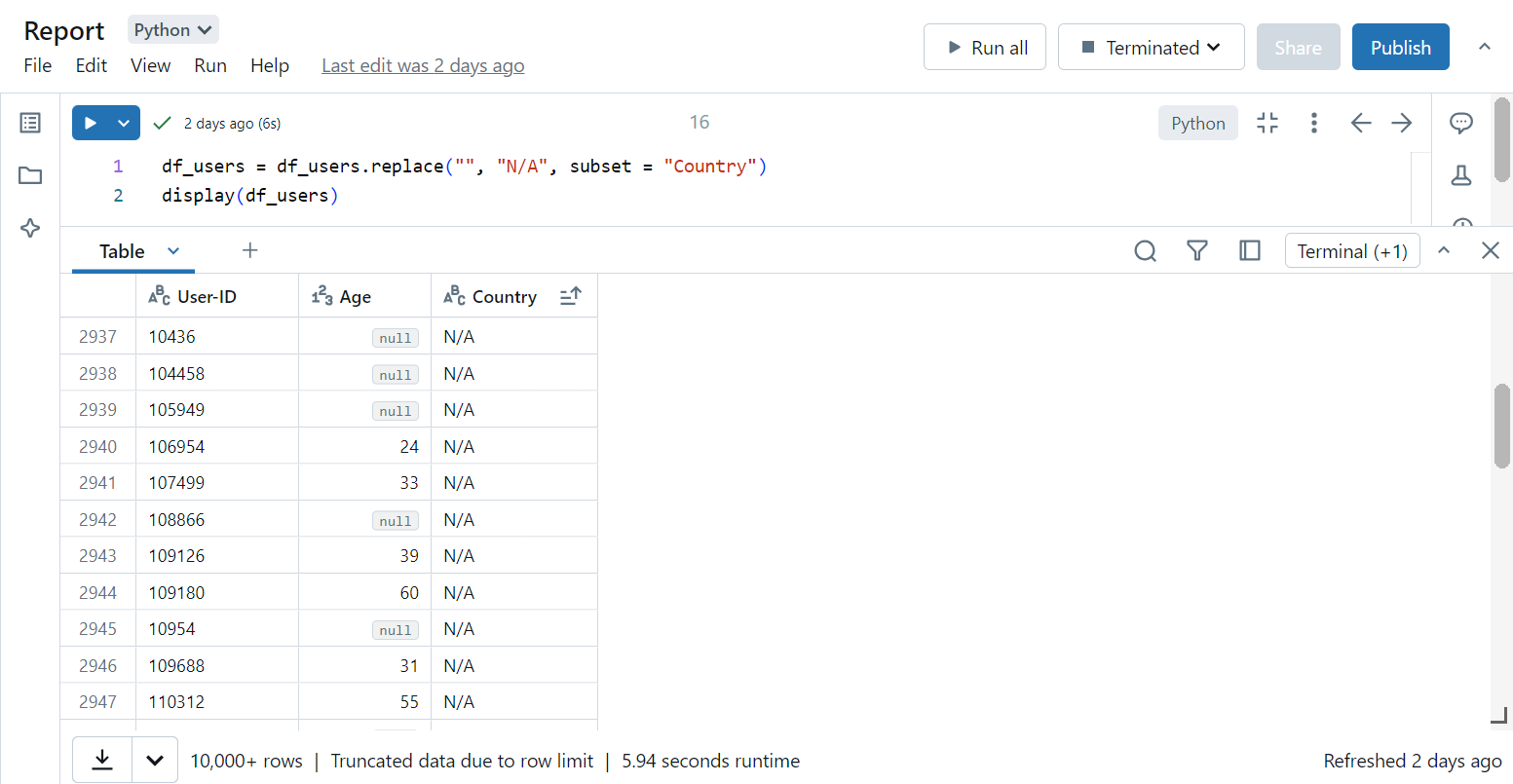
Hình 3.4‑16. Tách cột Location

##### Chuẩn hóa giá trị Null

Trong cột Country sẽ có những ô trống không có giá trị, thay vào đó chúng ta sẽ định dạng các ô trống đó trở thành một giá trị nào đó đồng bộ. Ở đây là giá trị ‘N/A’.

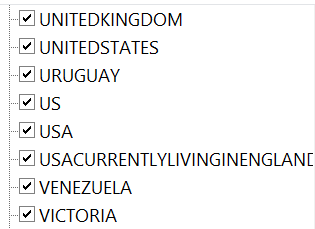


Hình 3.4‑17. Trước khi xử lý Null Country bảng Users



Hình 3.4‑18. Xử lý Null Country bảng Users

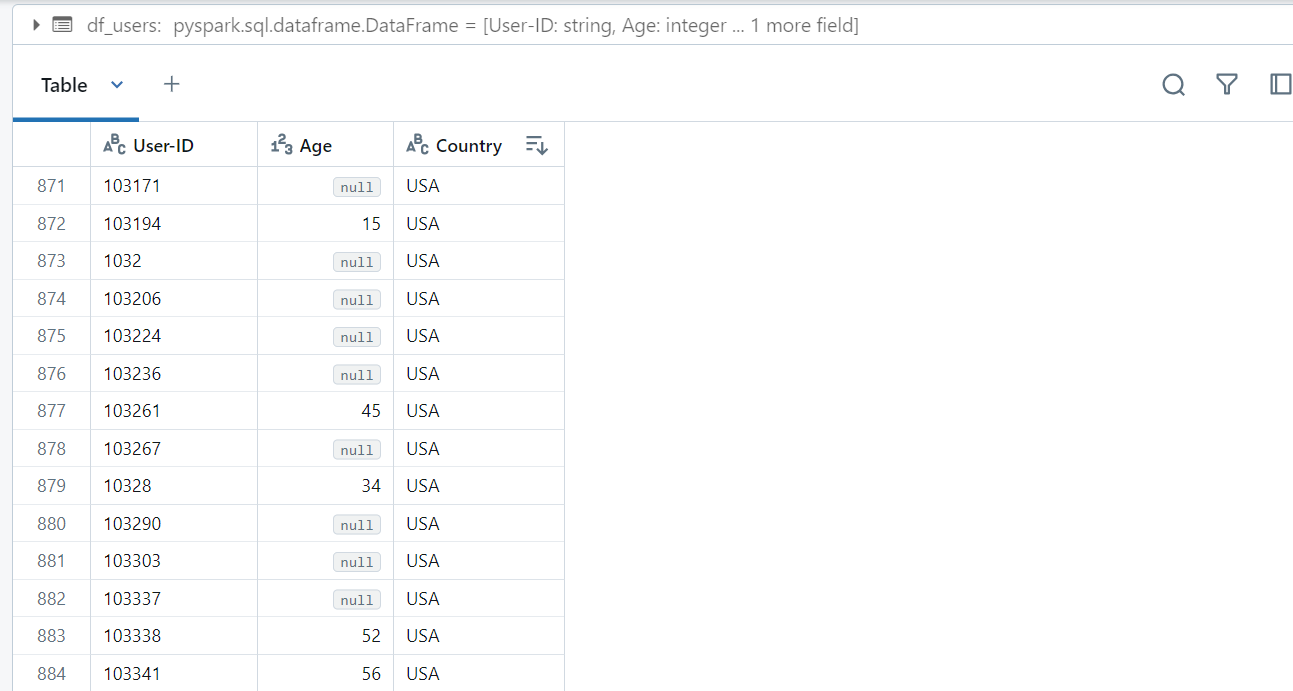
##### API chuẩn hóa cột Country

Vì sau xử lý các bước trên thì dữ liệu cột Country vẫn bị sai chính tả và không đồng bộ tên các quốc gia. Vì vậy nên có bước này, API vào một trang web khác để lấy dữ liệu tên các quốc gia và so sánh với tên các quốc gia có trong cột Country. Nếu như tên quốc gia ở trang web đã API có tên quốc gia tương tự với tên trong cột Country trên 80% thì lấy tên quốc gia trong trang web để thay vào cột Country.

Hình 3.4‑19. Cột Country chưa chuẩn hóa



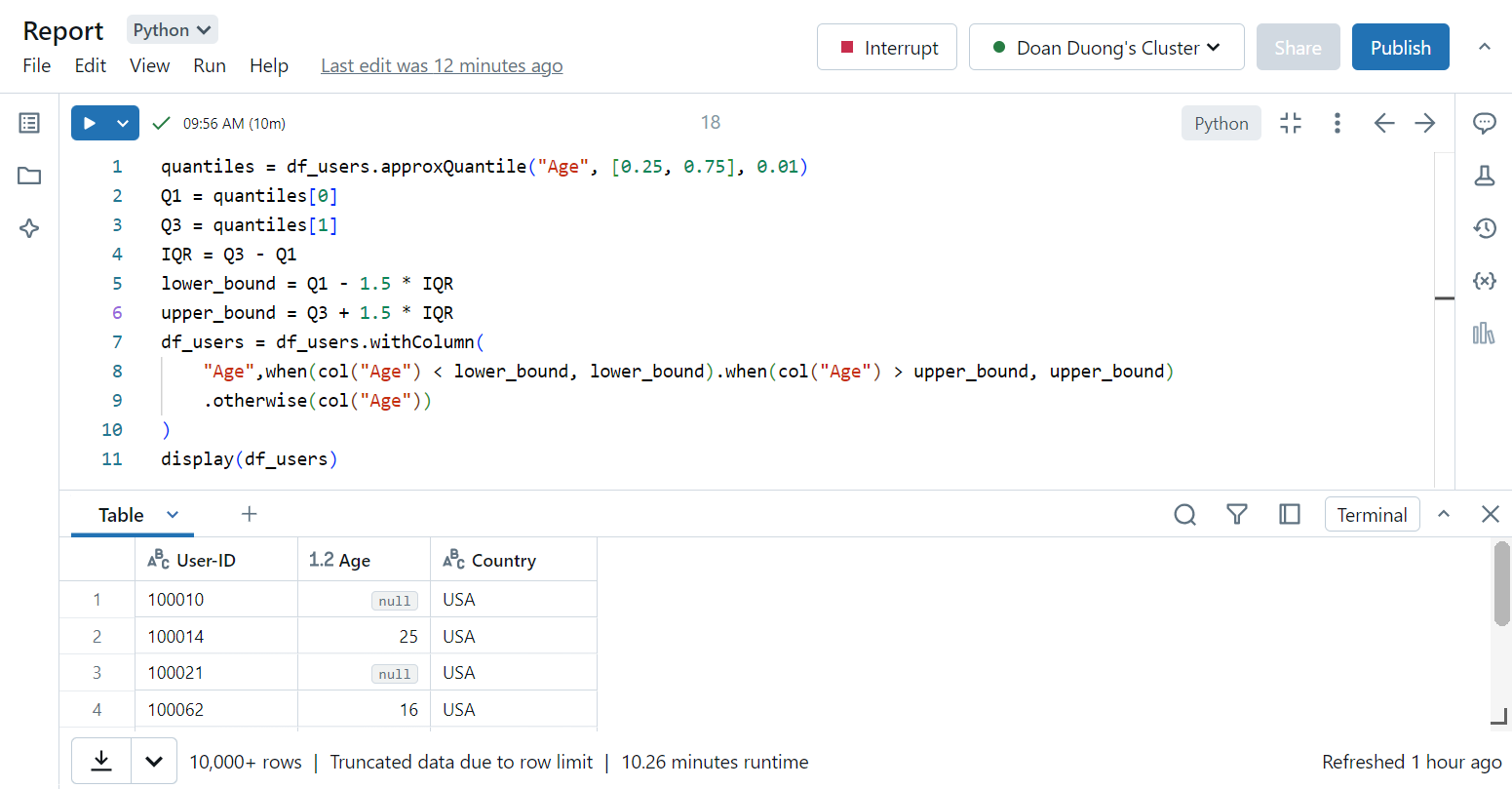
Hình 3.4‑20. Code chuẩn hóa cột Country



Hình 3.4‑21. Sau khi chuẩn hóa

##### Xử lý outlier

Vì sự tôn trọng dữ liệu chúng ta chỉ thay đổi giá trị của các giá trị outlier là min hoặc max của giá trị trong khoảng khả dụng chứ không xóa bỏ hoàn toàn các bản ghi đó.



Hình 3.4‑22. Xử lý outlier cột Age bảng Users

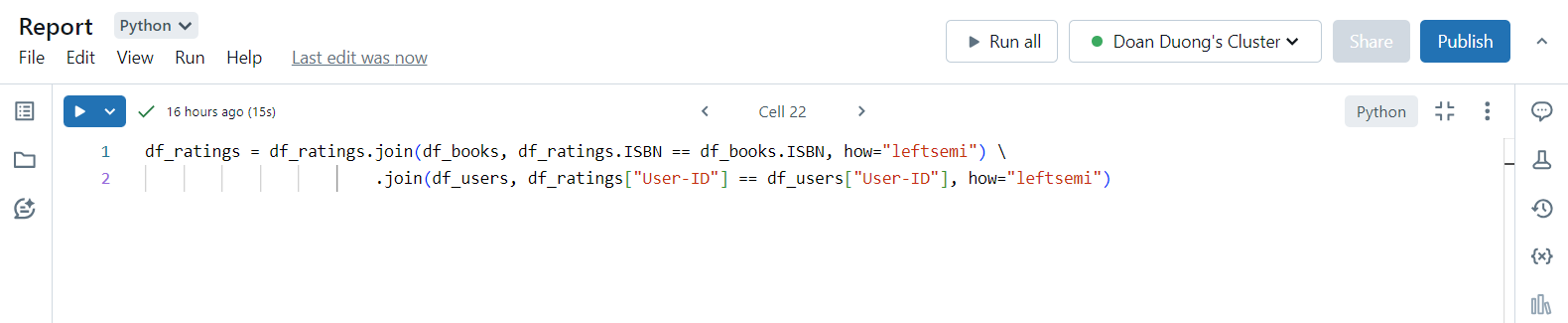
##### Điền dữ liệu Null cho cột Age

Dữ liệu cột Age sau khi xử lý outlier thì vẫn còn các bản ghi bị trống vậy thì chúng ta sẽ đến bước này để điền các dữ liệu trống đó bằng các giá trị tuổi trung bình theo từng quốc gia. Như vậy sẽ đỡ bị mất và trống dữ liệu hơn và cũng đáng tin cậy khi sử dụng độ tuổi để phân tích.



##### Ràng buộc quan hệ cho cơ sở dữ liệu

Để loại bỏ những đánh giá của các quyển sách không thuộc bảng Books và các đánh giá của ID người dùng không thuộc bảng Users thì chúng ta ràng buộc lại quan hệ bằng phép nối join leftsemi.

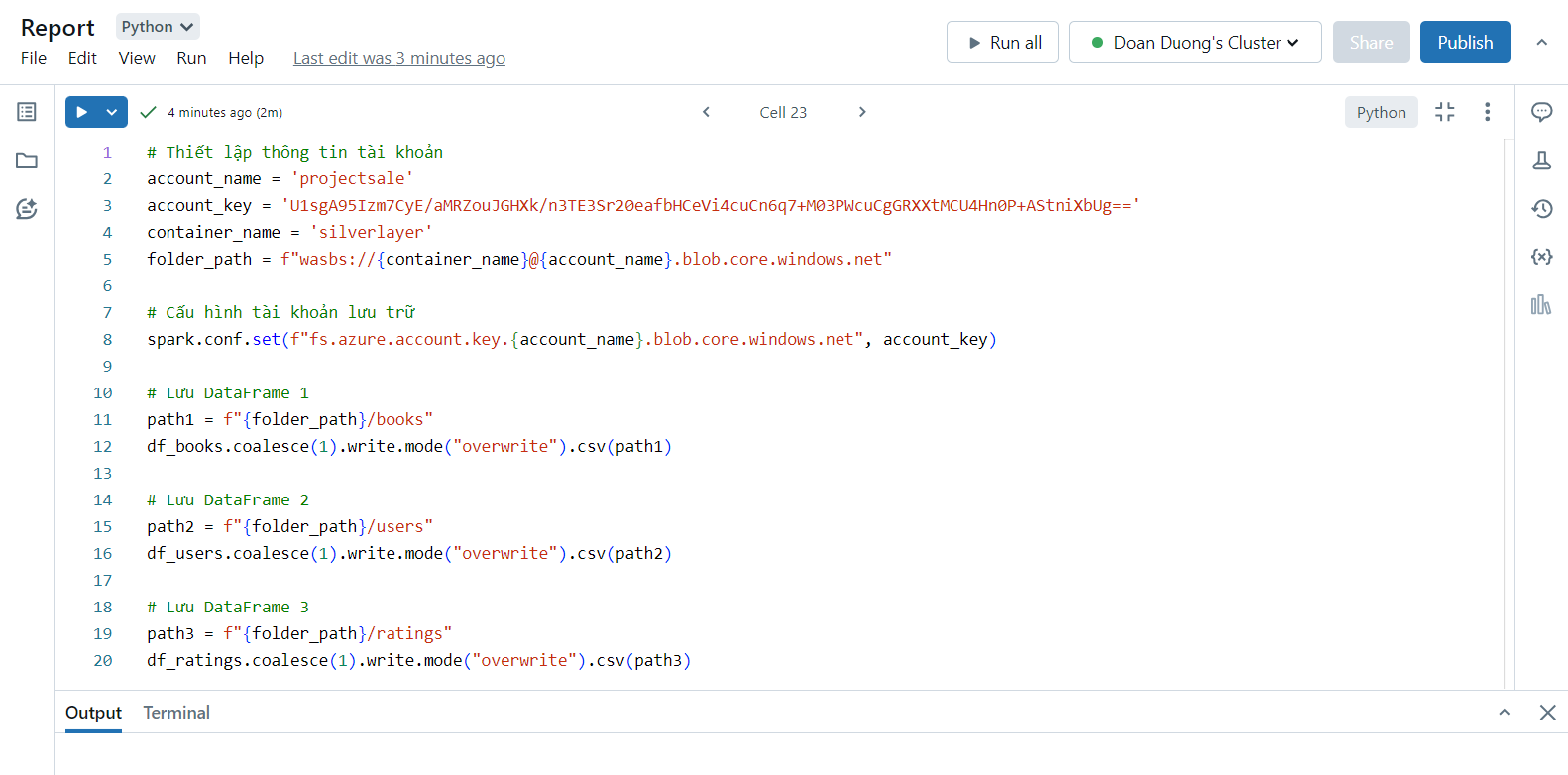


Hình 3.4‑23. Ràng buộc dữ liệu

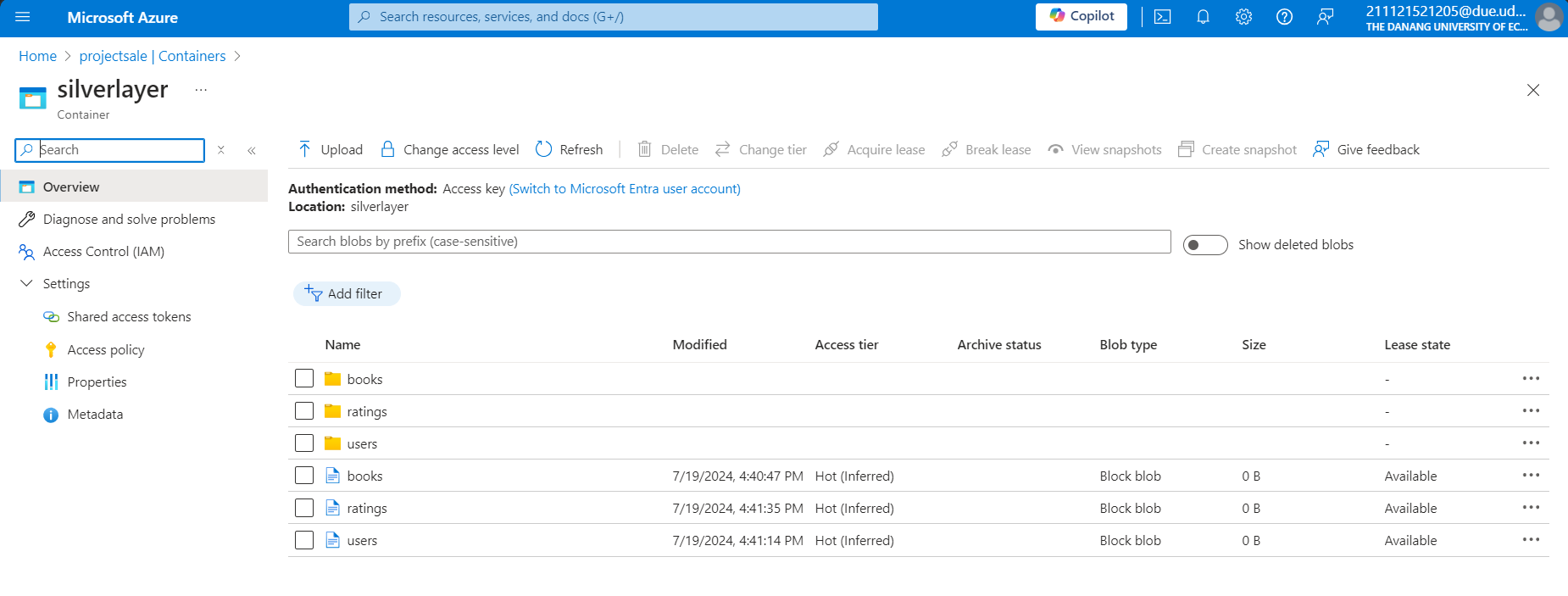
#### Lưu trữ dữ liệu Silver layer bằng Azure Blob Storage

Đảm bảo dữ liệu Silver Layer luôn sẵn sàng và an toàn, kể cả khi có sự cố xảy ra và nếu như có sự cố xảy ra với dữ liệu xử lý sau đó thì vẫn có thể lấy lại dữ liệu lúc này để xử lý.

Azure Blob Storage tương thích với nhiều công cụ và dịch vụ dữ liệu trong hệ sinh thái Azure và ngoài Azure, bao gồm Apache Hadoop, Apache Spark, và các công cụ ETL/ELT khác. Giúp dễ dàng tích hợp và sử dụng dữ liệu Silver Layer trong nhiều ứng dụng và trang web khác.



Hình 3.4‑24. Liên kết và lưu vào Azure Blob Storage

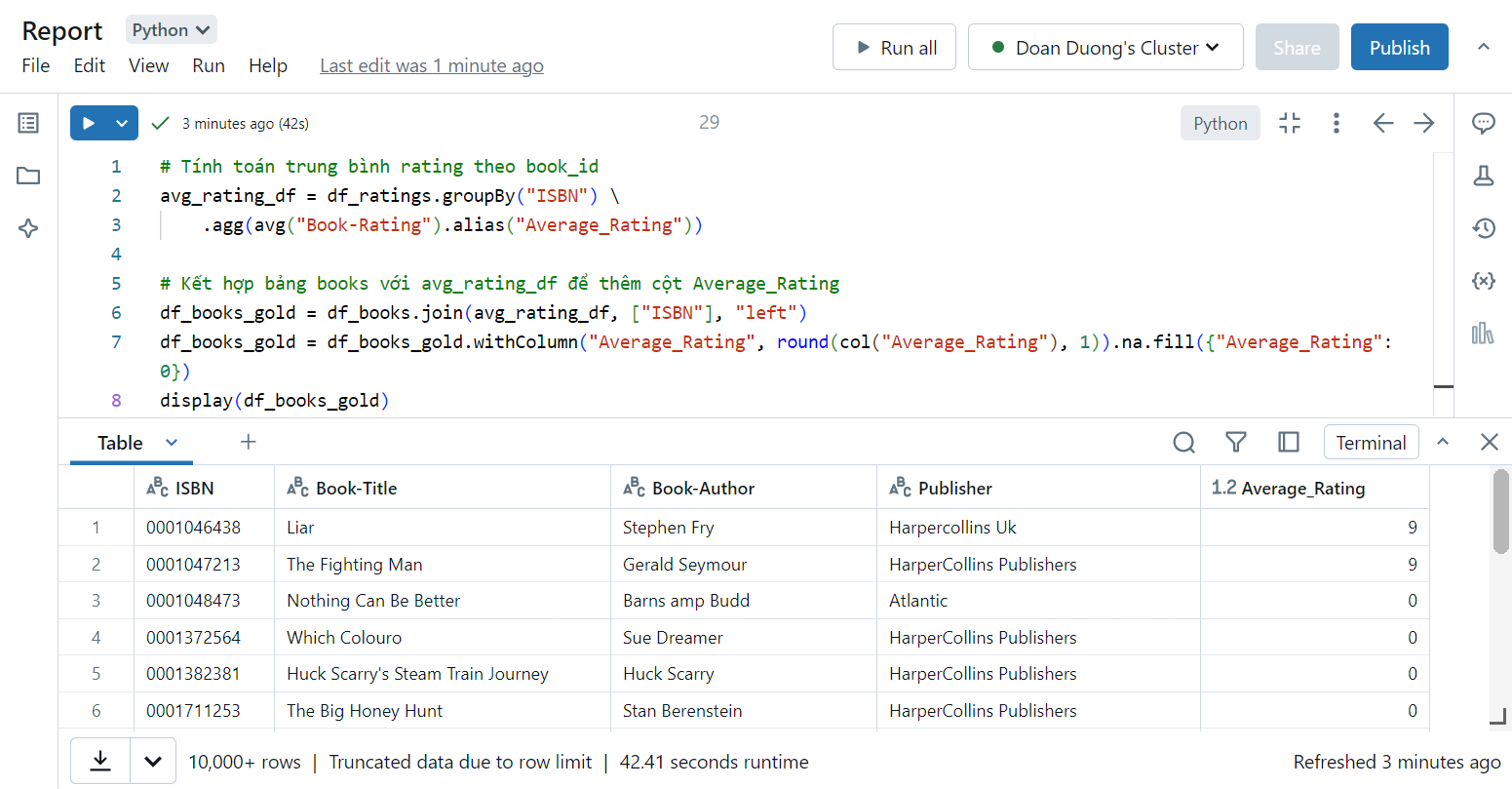


Hình 3.4‑25. Kết quả đưa lên Azure Blob Storage

#### Xử lý dữ liệu đưa vào Gold layer trên Databricks

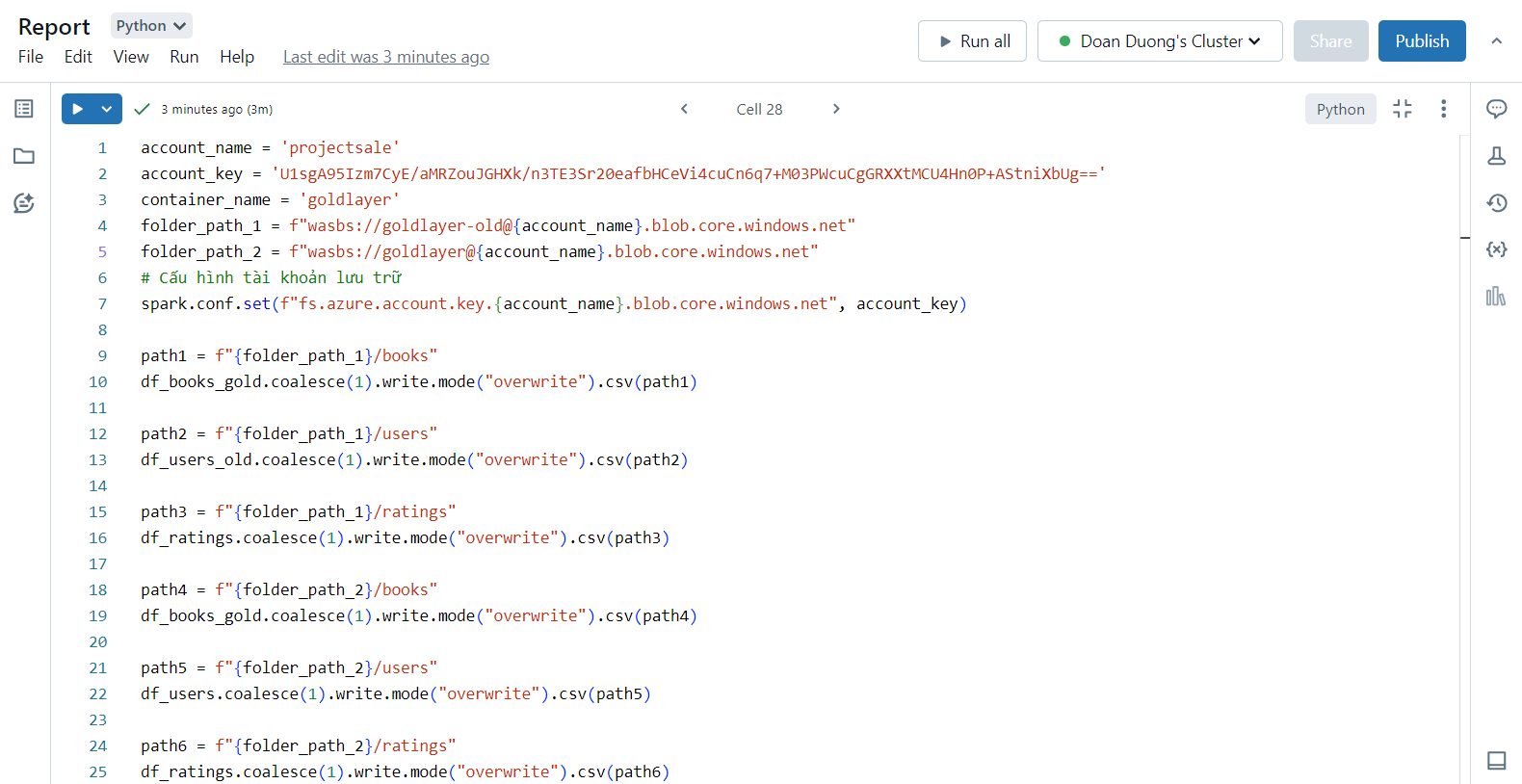
Sau khi lưu trữ dữ liệu Silver layer vào Azure Blob Storage thì chúng ta sẽ tiếp tục xử lý và lưu dữ liệu vào Gold layer. Vì yêu cầu của khách hàng là phân tích hành vi người đọc để có thể xuất bản những quyển sách bán chạy theo từng vùng, độ tuổi, … nên chúng ta sẽ thêm các cột cần thiết như trung bình đánh giá sách để có thể dễ dàng đánh giá cũng như tiện hơn trong việc đưa ra quyết định.

Ở bảng Books, để có thêm thông tin trung bình đánh giá chúng ra tính bằng cách trung bình các điểm đánh giá ở bảng Ratings và ghép cột Average\_Rating lại vào bảng Books.

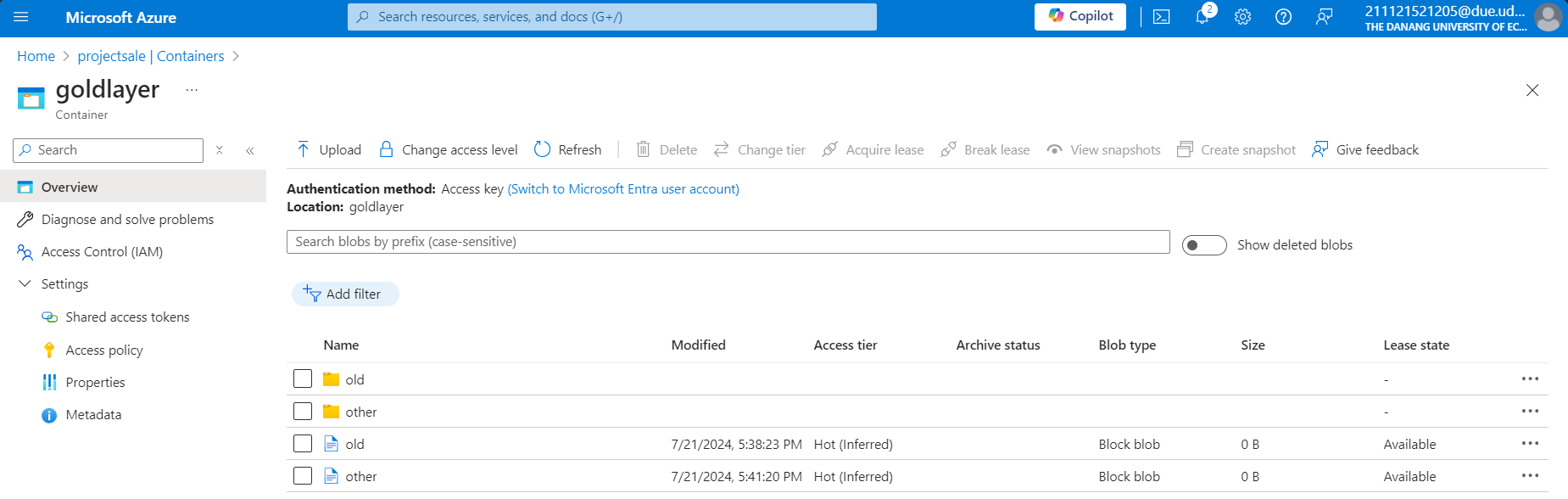


Hình 3.4‑26. Tạo thêm cột Average\_Ratings

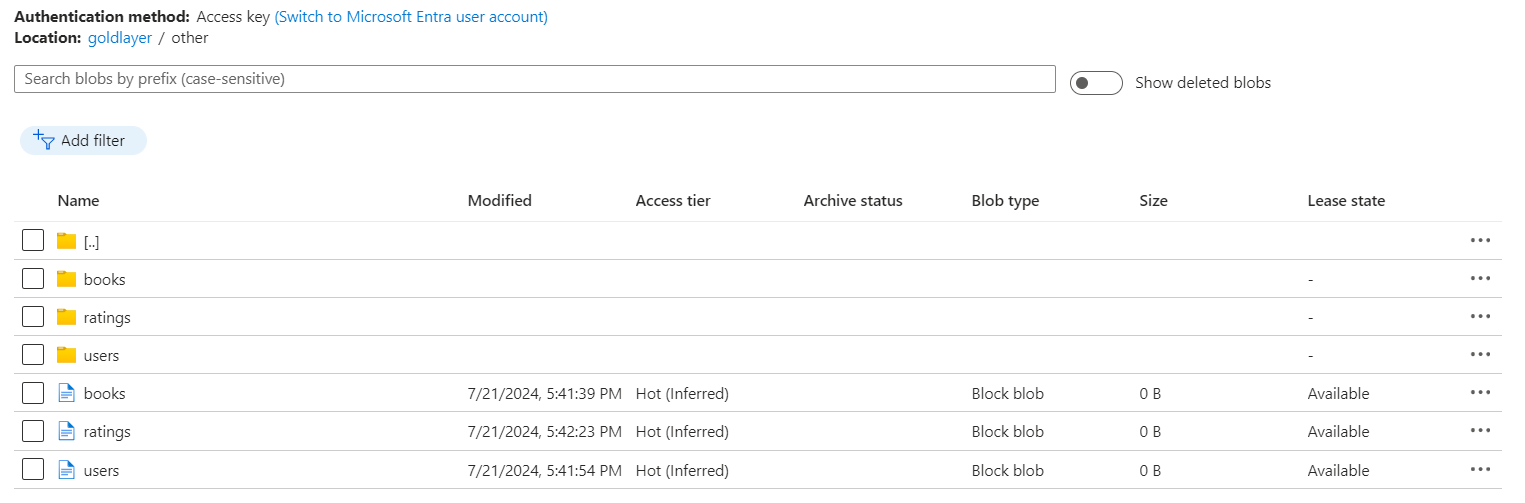
#### Lưu trữ dữ liệu Gold layer bằng Azure Blob Storage



Hình 3.4‑27. Kết nối và lưu trữ Gold layer lên Blob Storage



Hình 3.4‑28. Kết quả lưu trữ Gold layer

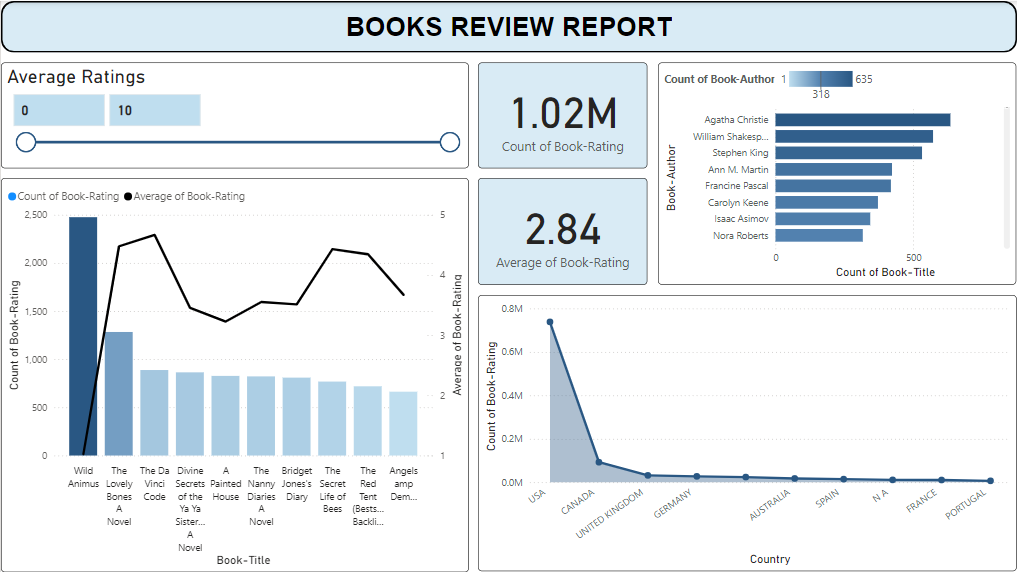


Hình 3.4‑29. Kết quả lưu trữ Gold layer

#### Phân tích bằng Power BI

Là một người truyền đạt dữ liệu thì trong Dashboard này chúng ta có thể thấy được các quyển sách được đánh giá nhiều nhất cũng như trung bình đánh giá, tác giả của các quyển sách đó cũng như những quốc gia có lượt đánh giá nhiều nhất.

Từ biểu đồ, chúng ta có thể rút ra những thông tin hữu ích để hỗ trợ quyết định của nhà xuất bản sách. Đầu tiên, việc xuất bản top 50 quyển sách có đánh giá cao nhất là một chiến lược tiềm năng để thu hút người đọc và tăng doanh thu. Thứ hai, tập trung vào các thị trường tiềm năng như Mỹ và Canada, nơi có số lượng độc giả và đánh giá sách cao, sẽ giúp mở rộng thị trường và tăng doanh số bán hàng. Cuối cùng, duy trì mối quan hệ với các tác giả có nhiều quyển sách được đánh giá trung bình đến cao là điều cần thiết. Nếu họ ra mắt sách mới, nhà xuất bản của chúng ta sẽ là đơn vị đầu tiên hợp tác, giúp nắm bắt thị trường kịp thời và tăng lợi nhuận cho công ty.



Hình 3.4‑30. Dashboard

### Kết luận

Chương này tập trung vào khảo sát yêu cầu và phân tích dữ liệu, giới thiệu các bảng dữ liệu như Books, Ratings, và Users, cùng với mối quan hệ giữa chúng. Tôi đã trình bày các công cụ cần dùng như DataBricks (Pyspark), Azure Blob Storage, và Power BI. Ngoài ra, tôi cũng mô tả quy trình xử lý dữ liệu thô, lưu trữ dữ liệu Silver layer và Gold layer. Qua chương này, người đọc có thể nắm bắt được quy trình thiết kế và quản lý cơ sở dữ liệu một cách chi tiết.

## TRIỂN KHAI THỰC NGHIỆM

### Kết luận

Trong chương cuối này, tôi đã tóm tắt các hoạt động thực hiện, đánh giá kết quả đạt được, và chia sẻ những bài học, kinh nghiệm rút ra từ quá trình thực tập. Tôi cũng đã nhận xét về môi trường làm việc tại doanh nghiệp và đề xuất những cải tiến cho tương lai. Kết luận tổng quát của chương này nhấn mạnh sự thành công của dự án thực tập và những kỹ năng, kiến thức đã thu được trong quá trình làm việc.

#### Tóm tắt các hoạt động đã thực hiện

Trong dự án thực tập tại Công ty TMA Solutions Bình Định, tôi đã tham gia vào việc xử lý và phân tích dữ liệu để hỗ trợ quyết định kinh doanh cho một nhà xuất bản sách. Dự án bao gồm ba file dữ liệu chính: Books, Ratings, và Users.

Xử lý dữ liệu: Tôi đã thực hiện các bước tiền xử lý dữ liệu như làm sạch dữ liệu, xử lý các giá trị thiếu và chuẩn hóa dữ liệu từ ba file.

Phân tích dữ liệu: Tôi đã thực hiện các phân tích thống kê và trực quan hóa dữ liệu để cung cấp cái nhìn sâu sắc về các xu hướng và mối quan hệ giữa các yếu tố. Điều này bao gồm phân tích các đánh giá sách, sự phân phối của các loại sách, và thói quen đọc của người dùng.

#### Đánh giá kết quả đạt được

Mục tiêu hoàn thành: Tôi đã hoàn thành việc xây dựng các mô hình phân tích dữ liệu để giúp nhà xuất bản đưa ra các quyết định kinh doanh dựa trên dữ liệu từ ba file. Cụ thể, tôi đã cung cấp các báo cáo và bảng điều khiển phân tích để hiểu rõ hơn về hiệu suất sách, sở thích của người đọc, và mối liên hệ giữa các yếu tố.

#### Những bài học và kinh nghiệm rút ra

Kỹ năng và kiến thức: Tôi đã cải thiện kỹ năng phân tích dữ liệu và trực quan hóa dữ liệu, cũng như học được cách sử dụng các công cụ xử lý như Python, Spark và các thư viện phân tích dữ liệu như Pandas và Numpy và công cụ trực quan hóa dữ liệu Power BI.

#### Nhận xét về môi trường làm việc

Trải nghiệm làm việc: Môi trường làm việc tại Công ty rất tích cực và hỗ trợ. Tôi đã nhận được sự hướng dẫn và phản hồi hữu ích từ các chuyên gia trong lĩnh vực phân tích dữ liệu và quản lý dự án.

Phát triển cá nhân: Kỳ thực tập đã giúp tôi nâng cao kỹ năng phân tích dữ liệu và khả năng đưa ra quyết định dựa trên dữ liệu thực tế. Tôi cảm thấy rằng trải nghiệm này đã chuẩn bị cho tôi những kỹ năng cần thiết để làm việc hiệu quả trong ngành công nghiệp xuất bản sách.

#### Kết luận tổng quát

Kết luận lại, dự án xử lý và phân tích dữ liệu cho nhà xuất bản sách đã cung cấp cái nhìn sâu sắc về hiệu suất sách và thói quen của người đọc. Sự hỗ trợ từ đội ngũ và các công cụ phân tích đã giúp tôi đạt được kết quả mong muốn và chuẩn bị tốt cho sự nghiệp phân tích dữ liệu trong tương lai. Tôi tin rằng những kỹ năng và kiến thức tích lũy được từ dự án này sẽ là nền tảng vững chắc cho công việc của tôi trong ngành xuất bản sách.

### Đề xuất

Cải tiến quy trình: Đề xuất việc triển khai thêm các công cụ phân tích dữ liệu nâng cao để theo dõi và phân tích hiệu suất của sách theo thời gian thực, cũng như cải thiện khả năng dự đoán xu hướng đọc sách.

Góp ý: Đề xuất xây dựng một hệ thống quản lý dữ liệu tích hợp để tổng hợp dữ liệu từ nhiều nguồn khác nhau, giúp việc phân tích và ra quyết định trở nên chính xác hơn.

# KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

* **Đạt được:**
* Kĩ năng lập trình: Python, Pyspark, SQL.
* Kĩ năng làm việc nhóm.
* Kĩ năng giao tiếp và giải quyết vấn đề.
* Khả năng sử dụng công cụ: Azure Blob Storage, Azure Data factory, Azure, DataBricks, Power BI.
* **Hạn chế:**
* Chưa hoàn thiện được kỹ năng tiếng anh.
* Kĩ năng thuyết trình còn yếu.
* **Hướng phát triển:**
* Trau dồi thêm tiếng anh và đạt được Toeic 550+ trong vòng 3 tháng tới.
* Cải thiện khả năng thuyết trình.

# TÀI LIỆU THAM KHẢO

|  |  |
| --- | --- |
| [1] | X. Hiep, "itguru," 23 11 2021. [Online]. Available: https://itguru.vn. [Accessed 10 6 2024]. |
| [2] | T. Vũ, "stringee," 30 5 2023. [Online]. Available: https://stringee.com. [Accessed 30 6 2024]. |

# CHECK LIST CỦA BÁO CÁO

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Nội dung công việc** | **Có** | **Không** | **Ghi chú** |
| 1 | Báo cáo được trình bày (định dạng) đúng với yêu cầu. | x |  |  |
| 2 | Báo cáo có số lượng trang đáp ứng đúng yêu cầu (30-50 trang) | x |  |  |
| 3 | Báo cáo trình bày được phần mở đầu bao gồm: Mục tiêu, Phạm vi và đối tượng, kết cấu … |  | x |  |
| 4 | Báo cáo trình bày về công ty, vị trí việc làm (công việc đó làm gì, kiến thức và kỹ năng cần thiết là gì, con đường phát triển sự nghiệp (career path)), cơ sở lý thuyết phù hợp với nội dung của đề tài (Tối đa 10-12 trang) | x |  |  |
| 5 | Báo cáo có sản phẩm cụ thể phù hợp với mục tiêu đặt ra của đề tài | x |  |  |
| 6 | Báo cáo có phần kết luận và hướng phát triển của đề tài | x |  |  |

# PHỤ LỤC