TRƯỜNG ĐẠI HỌC KINH TẾ

**KHOA THỐNG KÊ – TIN HỌC**

–––––––––––––––––––––––––––––––



**BÁO CÁO THỰC TẬP NGHỀ NGHIỆP**

**NGÀNH HỆ THỐNG THÔNG TIN QUẢN LÝ**

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

**TRỰC QUAN HÓA DỮ LIỆU ĐƯA RA XU HƯỚNG PHÁT TRIỂN CỦA NGÀNH CÔNG NGHIỆP ĐIỆN ẢNH HOLLYWOOD**

Sinh viên thực hiện : Hồ Thị Hồng Thắm

Nguyễn Lê Hoài Long

Phạm Tấn Thiên

Lớp : 45K21.1

Đơn vị thực tập : TMA Solutions Bình Định

Cán bộ hướng dẫn : Nguyễn Bá Văn

Nguyễn Khương Duy

Nguyễn Tấn Vỹ

Giảng viên hướng dẫn : ThS. Cao Thị Nhâm

**Đà Nẵng, 8/2022**

**NHẬN XÉT CỦA ĐƠN VỊ THỰC TẬP**

Họ và tên sinh viên: Hồ Thị Hồng Thắm

Lớp: 45K21.1

Khoa: Thống kê – Tin học

Trường: Đại học Kinh tế - Đại học Đà Nẵng

Thực tập từ ngày: 06/06/2022 đến ngày: 12/08/2022

Tại: TMA Solutions Bình Định

Địa chỉ: 12 Đại lộ Khoa học, Thung lũng Sáng tạo, Ghềnh Ráng, Tp. Qui Nhơn, Bình Định.

Sau quá trình thực tập tại đơn vị của sinh viên, chúng tôi có một số nhận xét, đánh giá như sau:

**1. Về thái độ, ý thức, đạo đức, kỷ luật**

**2. Kiến thức chuyên môn**

**3. Khả năng hòa nhập và thích nghi với công việc**

**4. Trách nhiệm, sáng tạo trong công việc**

**5. Các nhận xét khác**

**Đánh giá chung:**

**Điểm:**

Bình Định, ngày .......tháng ......năm 2022

**Xác nhận của đơn vị thực tập**

**NHẬN XÉT CỦA ĐƠN VỊ THỰC TẬP**

Họ và tên sinh viên: Nguyễn Lê Hoài Long

Lớp: 45K21.1

Khoa: Thống kê – Tin học

Trường: Đại học Kinh tế - Đại học Đà Nẵng

Thực tập từ ngày: 06/06/2022 đến ngày: 12/08/2022

Tại: TMA Solutions Bình Định

Địa chỉ: 12 Đại lộ Khoa học, Thung lũng Sáng tạo, Ghềnh Ráng, Tp. Qui Nhơn, Bình Định.

Sau quá trình thực tập tại đơn vị của sinh viên, chúng tôi có một số nhận xét, đánh giá như sau:

**1. Về thái độ, ý thức, đạo đức, kỷ luật**

**2. Kiến thức chuyên môn**

**3. Khả năng hòa nhập và thích nghi với công việc**

**4. Trách nhiệm, sáng tạo trong công việc**

**5. Các nhận xét khác**

**Đánh giá chung:**

**Điểm:**

Bình Định, ngày .......tháng ......năm 2022

**Xác nhận của đơn vị thực tập**

**NHẬN XÉT CỦA ĐƠN VỊ THỰC TẬP**

Họ và tên sinh viên: Phạm Tấn Thiên

Lớp: 45K21.1

Khoa: Thống kê – Tin học

Trường: Đại học Kinh tế - Đại học Đà Nẵng

Thực tập từ ngày: 06/06/2022 đến ngày: 12/08/2022

Tại: TMA Solutions Bình Định

Địa chỉ: 12 Đại lộ Khoa học, Thung lũng Sáng tạo, Ghềnh Ráng, Tp. Qui Nhơn, Bình Định.

Sau quá trình thực tập tại đơn vị của sinh viên, chúng tôi có một số nhận xét, đánh giá như sau:

**1. Về thái độ, ý thức, đạo đức, kỷ luật**

**2. Kiến thức chuyên môn**

**3. Khả năng hòa nhập và thích nghi với công việc**

**4. Trách nhiệm, sáng tạo trong công việc**

**5. Các nhận xét khác**

**Đánh giá chung:**

**Điểm:**

Bình Định, ngày .......tháng ......năm 2022

**Xác nhận của đơn vị thực tập**

# LỜI CẢM ƠN

Lời cảm ơn đầu tiên nhóm em xin gửi đến quý thầy, cô giáo trong khoa Thống kê – Tin học trường Đại học Kinh tế - Đại học Đà Nẵng lời cảm ơn chân thành vì đã tận tâm giảng dạy và truyền đạt những kiến thức, kinh nghiệm quý báu cho nhóm. Đặc biệt nhóm xin gửi lời cảm ơn sâu sắc nhất đến cô Cao Thị Nhâm – người đã tận tình hướng dẫn, giúp đỡ nhóm hoàn thành chuyên đề báo cáo thực tập này.

Nhóm xin chân thành cảm ơn ban lãnh đạo cùng với các phòng ban, các cô chú, anh chị trong Công ty TMA Solutions Bình Định – đơn vị đã tiếp nhận và đã tạo điều kiện thuận lợi cho nhóm tiếp cận thực tế, nắm bắt được các quy trình làm việc.

Cuối cùng nhóm xin cảm ơn các anh Nguyễn Khương Duy, Nguyễn Bá Văn, Nguyễn Tấn Vỹ - là những Mentors hướng dẫn trực tiếp các thành viên trong nhóm tại công ty TMA. Các anh đã tận tình giúp đỡ, cung cấp những tài liệu cho chúng em để hoàn thành bài báo cáo trong quá trình thực tập mảng Data Engineer tại công ty.

Vì thời gian và kiến thức còn hạn hẹp nên bài báo cáo không thể tránh khỏi những thiếu sót, rất mong sự góp ý của công ty, quý thầy cô để các thành viên trong nhóm rút kinh nghiệm và hoàn thành tốt hơn.

Nhóm chúng em xin chân thành cảm ơn!

# LỜI CAM ĐOAN

Nhóm chúng em xin cam đoan dự án “Thu thập dữ liệu, xử lý và đánh giá ngành công nghiệp điện ảnh” là kết quả nghiên cứu của nhóm dưới sự hướng dẫn của giảng viên Cao Thị Nhâm và các anh mentors Nguyễn Khương Duy, Nguyễn Bá Văn, Nguyễn Tấn Vỹ. Ngoài ra không có bất cứ sự sao chép nào của người khác.

Đề tài, nội dung báo cáo thực tập là sản phẩm mà nhóm em đã nỗ lực nghiên cứu trong suốt quá trình thực tập tại công ty TMA Solutions Bình Định. Các dữ liệu trong bài báo cáo là hoàn toàn trung thực. Chúng em xin chịu hoàn toàn trách nhiệm, kỷ luật của nhà trường đề ra nếu có vấn đề nào xảy ra.

Quy định:

Soạn thảo trên trang A4 (trang dọc, lề trái: 3.5cm; trên, phải, dưới: 2.5 cm), Font Times New Roman, canh đều 2 bên, size 13, cách dòng 1.5, cách đoạn trên 6pt, cách đoạn dưới 3pt, hàng đâu tiên lùi vào 1.27 cm; hình và bảng soạn theo caption, chèn trích dẫn chéo (Cross-reference) cho bảng và hình; các danh mục hình, bảng, mục lục làm tự động; đánh số trang như file mẫu (bìa không có số trang, danh mục + mục lục số trang theo i, ii, iii…, nội dung chính theo 1,2,3…)

# MỤC LỤC

[NHẬN XÉT CỦA DOANH NGHIỆP THỰC TẬP ii](#_Toc98339732)

[LỜI CẢM ƠN iii](#_Toc98339733)

[LỜI CAM ĐOAN iv](#_Toc98339734)

[MỤC LỤC vi](#_Toc98339735)

[DANH MỤC HÌNH ẢNH viii](#_Toc98339736)

[DANH MỤC BẢNG BIỂU ix](#_Toc98339737)

[DANH MỤC CÁC TỪ VIẾT TẮT x](#_Toc98339738)

[LỜI MỞ ĐẦU 1](#_Toc98339739)

[CHƯƠNG 1. TỔNG QUAN … 2](#_Toc98339740)

[1.1. Mục 1.1 2](#_Toc98339741)

[1.1.1. Mục 1.1.1 2](#_Toc98339742)

[1.1.2. Mục 1.1.2 2](#_Toc98339743)

[1.2. Mục 1.2 2](#_Toc98339744)

[CHƯƠNG 2. LÝ THUYẾT 3](#_Toc98339745)

[2.1. Mục 2.1 3](#_Toc98339746)

[2.1.1. Mục 2.1.1 3](#_Toc98339747)

[2.1.2. Mục 2.1.2 3](#_Toc98339748)

[2.2. Mục 2.2 3](#_Toc98339749)

[CHƯƠNG 3. TRIỂN KHAI … 4](#_Toc98339750)

[3.1. Mục 3.1 4](#_Toc98339751)

[3.1.1. Mục 3.1.1 4](#_Toc98339752)

[3.1.2. Mục 3.1.2 4](#_Toc98339753)

[3.2. Mục 3.2 4](#_Toc98339754)

[CHƯƠNG 4. KẾT QUẢ 5](#_Toc98339755)

[4.1. Mục 4.1… 5](#_Toc98339756)

[4.2. Mục 4.2… 5](#_Toc98339757)

[KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 6](#_Toc98339758)

[TÀI LIỆU THAM KHẢO 7](#_Toc98339759)

[PHỤ LỤC 8](#_Toc98339760)

*(Mục lục này chỉ là ví dụ)*

# DANH MỤC HÌNH ẢNH

[Hình 1.1 Ngôn ngữ lập trình Python 2](#_Toc98336120)

[Hình 4.1 Kiến trúc của mô hình 5](#_Toc98336121)

# DANH MỤC BẢNG BIỂU

[Bảng 4.1 Kiến trúc 5](#_Toc74235471)

# DANH MỤC CÁC TỪ VIẾT TẮT

**AI** : Artificial Intelligence

**…**

# LỜI MỞ ĐẦU

1. **Mục tiêu nghiên cứu của đề tài**

* Đề tài này nghiên cứu …

1. **Nhiệm vụ của đề tài**

* Nghiên cứu ...
* ...

1. **Phương pháp nghiên cứu**

* ...

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

* ...

1. **Kết cấu của đề tài**

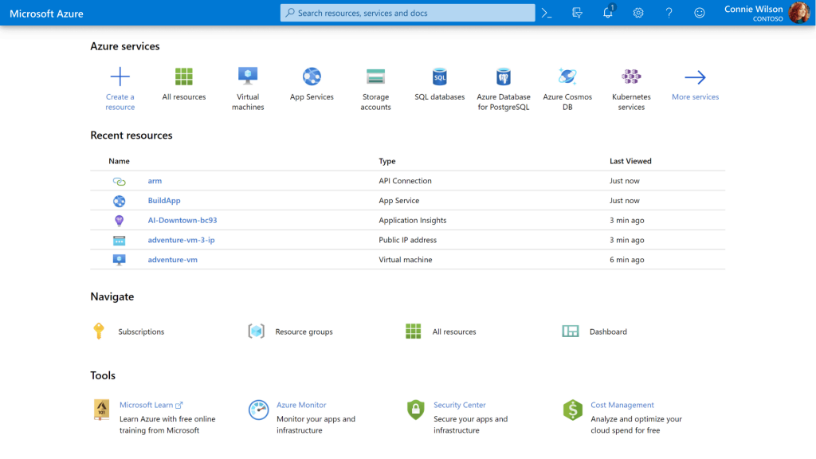
Đề tài được tổ chức gồm phần mở đầu, x chương nội dung và phần kết luận...

* Mở đầu
* **Chương 1**:
* **Chương 2**:
* **Chương 3**:
* **Chương 4**:
* Kết luận và hướng phát triển

# CƠ SỞ LÝ THUYẾT

## Microsoft Azure

### Microsoft Azure là gì?



Azure Microsoft là một nền tảng điện toán đám mây (cloud computing platform) và một cổng trực tuyến (online portal) cho phép truy cập và quản lý các dịch vụ (service) và tài nguyên (resource) đám mây do Microsoft cung cấp.

Những điều cần biết về Azure:

* Azure được phát hành vào ngày 1 tháng 2 năm 2010.
* Azure theo mô hình miễn phí để bắt đầu sử dụng và sau đó trả tiền cho mỗi lần sử dụng.
* 80% trong số 500 công ty trong danh sách Fortune sử dụng dịch vụ Azure cho nhu cầu điện toán đám mây.
* Azure hỗ trợ nhiều ngôn ngữ lập trình.
* Azure có số lượng trung tâm dữ liệu trên khắp thế giới.

### Cloud Computing (Điện toán đám mây)

1. Cloud Computing là gì?

Cloud Computing là việc cung cấp các dịch vụ điện toán hoàn toàn qua Internet. Hay nói đúng hơn là việc cung cấp tài nguyên phù hợp với nhu cầu người dùng hoàn toàn thông qua Internet.



1. Ưu điểm của Cloud Computing

* Khả năng mở rộng linh hoạt về quy mô.
* Khả năng phân bổ nguồn lực một cách nhanh chóng.
* Khả năng duy trì hoạt động trong các lỗi của bộ phận và dịch vụ.
* Khả năng phục hồi sau thảm họa. Chúng sao chép các tài nguyên tại các địa điểm khác nhau để có thể hoạt động ngay cả trong thời gian thảm họa.
* Khả năng giữ cho các tài nguyên có thể truy cập được trong một thời gian dài với thời gian ngừng hoạt động tối thiểu.

1. Các dạng Cloud Computing

**Public Cloud:** Là loại hình được sở hữu và vận hành bởi nhà cung cấp dịch vụ trung gian. Với tất cả phần cứng, phần mềm và cơ sở hạ tầng hỗ trợ khác đều do nhà cung cấp sở hữu và quản lý.

**Private Cloud:** Là một dạng tài nguyên được sử dụng riêng cho một tổ chức hoặc doanh nghiệp nào đó. Lúc này các tài nguyên của hệ thống công nghệ thông tin trong doanh nghiệp đều được chia sẻ về hệ thống máy chủ tính toán, giúp doanh nghiệp tận dụng được các tài nguyên này một cách hiệu quả và bảo mật nhất, cũng như tối ưu được nguồn tài chính, giảm thiểu rủi ro đầu tư cho chi phí cố định.

**Hybrid Cloud:** Là sự kết hợp giữa cả hai hình thức trên. Chúng được ràng buộc với nhau bằng công nghệ để đảm bảo sự riêng tư. Lựa chọn này sẽ giúp cho doanh nghiệp có nhiều lựa chọn hơn, giúp tối ưu hóa cơ sở hạ tầng hiện tại.

### Understand Azure Core Services

##### Vùng Azure: Có các khu vực trên toàn thế giới có thể được sử dụng để lưu trữ tài nguyên.

##### Vùng khả dụng: Là những địa điểm riêng biệt về mặt vật lý trong một vùng Azure. Mỗi vùng khả dụng được tạo thành từ một hoặc nhiều trung tâm dữ liệu. Có thể sử dụng vùng sẵn sàng để chạy các ứng dụng quan trọng với tính khả dụng cao và sao chép độ trễ thấp.

##### Mạng ảo: Bạn có thể xây dựng mạng ảo trong Azure. Mỗi mạng ảo trong Azure cần được gán cho một không gian địa chỉ. Một số mạng con có không gian địa chỉ là một tập con của mạng ảo.

##### Ứng dụng Web:

* Mục đích
  + Lưu trữ một ứng dụng web.
  + Lưu trữ một ứng dụng web - Nhiều ứng dụng/ Mở rộng quy mô.
  + Cài đặt các bản cập nhật bảo mật.
* Tính năng
* Net, .Net Core, Java, Ruby, Node.JS, PHP, Python.
* Các phục vụ cơ bản có thể là Windows hoặc Linux.
* DevOps-Triển khai liên tục.
* Triển khai Blue-Green bằng cách sử dụng các khe cắm dàn.
* Xác thực với các nhà cung cấp bên ngoài -Azure AD, Facebook, Google.

##### Windows Virtual Desktop:

**Host pool:** Bộ sưu tập các máy ảo Azure.

**Nhóm ứng dụng:** Làmột nhóm hợp lý các ứng dụng được cài đặt trên máy chủ phiên**.** Có hai loại nhóm ứng dụng:

* Ứng dụng từ xa: Tại đây người dùng truy cập các ứng dụng từ xa được chọn riêng lẻ. Sau đó, chúng được xuất bản lên nhóm ứng dụng.
* Máy tính để bàn: Tại đây người dùng truy cập toàn bộ máy tính để bàn.

**Workspace:** Không gian làm việc: Là cách phân nhóm hợp lý các nhóm ứng dụng trong Windows Virtual Desktop.

**End Users:** Tại đây bạn chỉ định người dùng vào các nhóm ứng dụng.

**Azure HDinsight:** Là bản phân phối đám mây của các thành phần dựa trên Hadoop. Dịch vụ này giúp xử lý lượng lớn dữ liệu dễ dàng hơn dựa trên các khung mã nguồn mở. Bạn có thể tạo các cụm Azure HDInsight cho nhu cầu xử lý hàng loạt.

**Azure Resource Manager Templates:** Là một mẫu JSON có chứa các tài nguyên mà bạn muốn tạo trong Azure. Bạn có thể gửi mẫu này đến Trình quản lý tài nguyên Azure. Trình quản lý tài nguyên Azure sau đó sẽ tạo tài nguyên dựa trên định nghĩa mẫu. Điều này tốt khi bạn cần tạo cùng một tập hợp tài nguyên - Kiểm tra môi trường. Sự khác biệt giữa các mẫu:

* ContentVersion: Là phiên bản của các mẫu
* Tài nguyên: Là phần chính khi bạn xác định các tài nguyên cần được triển khai như một phần của mẫu.
* Tham số: Là các giá trị có thể được cung cấp cho mẫu có thể được sử dụng để tùy chỉnh việc triển khai các tài nguyên.
* Biến: Là những giá trị có thể được sử dụng lại trong mẫu.
* Cognitive Services: Là một tập hợp các dịch vụ trong Azure giúp bạn xây dựng trí thông minh nhận thức trong các ứng dụng của mình. Dịch vụ nhận thức được chia thành năm loại chính: Vision, Speech, Language, Decision, Tìm kiếm.

##### Azure Bot Service: Toàn bộ ý tưởng về Bots - Giúp cung cấp trải nghiệm cho người dùng tương tác với người khác chứ không phải với máy tính. Bots là các chương trình tương tác đọc những gì một người đang cố gắng hỏi và sau đó dựa trên yêu cầu của người đó, bot cần trả lời người đó. Bot có thể giải thích những gì người dùng đang cố gắng nói và sau đó tạo ra một phản hồi thích hợp.

##### Azure DevOps: Là sự kết hợp của các khía cạnh văn hóa, thực hành và công cụ. Các khía cạnh này được sử dụng để tập hợp các nhóm lại với nhau và mục tiêu đó là cung cấp các ứng dụng và dịch vụ với tốc độ cao. Khi nói đến các công cụ, có nhiều công cụ khác nhau có sẵn trên thị trường để đạt được mục đích này:

* Azure Boards: Tại đây các nhóm có thể tạo và theo dõi các câu chuyện của người dùng, các mục tồn đọng, nhiệm vụ, tính năng và lỗi được liên kết với một dự án.
* Azure Repos: Là một bộ công cụ kiểm soát phiên bản giúp bạn quản lý mã của mình.
* Azure Pipelines: Có thể được sử dụng để tự động xây dựng và kiểm tra dự án mã của bạn.
* Kế hoạch thử nghiệm Azure: Bạn có thể quản lý thử nghiệm thủ công, bao gồm thử nghiệm chấp nhận của người dùng, thử nghiệm khám phá và phản hồi của người chia sẻ.
* Azure Artifacts: Bạn có thể tạo và chia sẻ nguồn cấp dữ liệu gói Maven, npm và NuGet từ các nguồn công khai và riêng tư với các nhóm ở mọi quy mô.

##### Azure DevOps – Understanding the lifecycle

Mọi dịch vụ trong Azure đều tuân theo vòng đời dịch vụ của chính nó. Hầu như tất cả các dịch vụ trong azure đều có ba giai đoạn:

* Private Preview: Các dịch vụ này chỉ có sẵn cho các loại khách hàng cụ thể.
* Public Preview: Tất cả các khách hàng đều có quyền truy cập vào các dịch vụ này. Loại dịch vụ này không kết thúc, nó không phù hợp với các công đoạn sản xuất; hơn nữa, dịch vụ dưới chế độ xem trước công khai không có SLA.
* General Availability (Tính sẵn có chung): Các dịch vụ này có thể truy cập được cho cả cộng đồng, hỗ trợ SLA và phù hợp với môi trường sản xuất.

##### Azure Logic Apps: Là dịch vụ được quản lí hoàn toàn và không có máy chủ. Bạn có thể xây dựng quy trình công việc và không cần phải lo lắng về cơ sở hạ tầng để lưu trữ quy trình công việc. Quy trình làm việc có thể tích hợp với các dịch vụ Azure khác nhau và các dịch vụ của bên thứ ba Bạn có thể xây dựng quy trình công việc thông qua Trình thiết kế trực quan.

##### Azure CosmosDB: Cung cấp 99,999% khả năng đọc và ghi, khả năng mở rộng quy mô từ hàng nghìn đến hàng trăm triệu yêu cầu/ giây. Đảm bảo độ trễ dưới 10 ms cho các lần đọc và ghi được lập chỉ mục ở phân vị thứ 99. Làm việc với các API khác nhau.

##### Azure SQL Database:

* SQL PaaS:
* Tính khả dụng 99,99%
* Sao lưu và phục hồi tích hợp
* Một số tính năng tối thiểu có thể không khả dụng
* Giảm chi phí cơ sở hạ tầng và bảo trì
* SQL IaaS:
* Bạn có toàn quyền kiểm soát công cụ SQL Server
* Dễ dàng di chuyển một phiên bản tại chỗ hiện có
* Bạn cần quản lý tính khả dụng

##### Azure Database for PostgreSQL: Là một nền tảng như một dịch vụ có sẵn cho PostgreSQL Community Edition. Có được tính khả dụng cao được tích hợp sẵn. Bạn cũng nhận được các bản sao lưu tự động và khôi phục tại thời điểm trong 35 ngày. Dịch vụ cơ bản chịu trách nhiệm duy trì phần cứng bên dưới và giữ cho công cụ cơ sở dữ liệu được cập nhật. Ở đây công ty không phải chịu chi phí duy trì cơ sở hạ tầng bên dưới.

## Azure Databricks

Đây là nền tảng phân tích dựa trên Apache Spark được tối ưu hóa cho đám mây Microsoft Azure. Với Azure Databricks, bạn có thể tạo các cụm Apache Spark được quản lý hoàn toàn. Bạn có thể động các cụm tỷ lệ tự động. Bạn cũng có thể tạo các cụm máy chủ và cũng có thể chia sẻ chúng giữa các nhóm và dễ dàng khám phá dữ liệu với sự trợ giúp của sổ ghi chép bằng R, Python, Scala và SQL. Bạn có thể sử dụng trang tổng quan tương tác để tạo báo cáo động. Nó có các khía cạnh bảo mật khi nói đến kiểm soát truy cập dựa trên vai trò và tích hợp với Azure Active Directory.

Apache Spark là một công cụ đa ngôn ngữ để thực thi kỹ thuật dữ liệu, khoa học dữ liệu và học máy trên các máy hoặc nút đơn.

Các tính năng chính:

* Dữ liệu hàng loạt/ Truyền trực tuyến: Hợp nhất việc xử lý dữ liệu của bạn theo lô và phát trực tuyến theo thời gian thực, sử dụng ngôn ngữ ưa thích của bạn: Python, SQL, Scala, Java hoặc R.
* Phân tích SQL: Thực thi các truy vấn ANSI SQL phân tán, nhanh chóng để lập bảng điều khiển và báo cáo đặc biệt. Chạy nhanh hơn hầu hết các kho dữ liệu.
* Khoa học dữ liệu ở quy mô: Thực thi phân tích dữ liệu khám phá (EDA) trên dữ liệu quy mô petabyte mà không cần phải lấy mẫu xuống.

## Data Engineer

Data Engineer hay còn gọi là Kỹ sư dữ liệu, người này được xem là người xây dựng, kiểm tra và duy trì kiến trúc tổng hợp, lưu trữ và xuất dữ liệu từ những app hay hệ thống được tạo bởi Software Engineer. Đồng thời, DE là người thực hiện cải biến các quy trình thiết lập dữ liệu để mô hình hóa, khai thác và sản xuất dữ liệu. Nói một cách ngắn gọn, DE là người thu thập, tổng hợp dữ liệu từ nhiều nguồn thành một nguồn duy nhất. Hệ thống được xây dựng từ DE dùng phục vụ cho các phòng ban nghiệp vụ, Data Analyst, Data Scientist, Business Intelligence phân tích dữ liệu để hỗ trợ cho việc ra quyết định.

Công việc chủ yếu của một DE bao gồm: Cấu trúc dữ liệu nâng cao (distributed computing), lập trình đồng thời (concurrent programming), kiến thức sâu về một số công cụ mới như: Hadoop, Spark, Kafka, Hive…, tạo ETL/ data pipelines.

R và Python chính là 2 ngôn ngữ phổ biến nhất đối với nhân viên DE. Làm việc với 2 ngôn ngữ này, họ sẽ thường xuyên sử dụng các packages như ggplot2 để thực hiện trực quan hóa dữ liệu trong R hoặc thao tác dữ liệu thư viện Pandas Python. Tại một số công ty, ngôn ngữ Scala cũng được chọn làm ngôn ngữ chính dùng để xử lý dữ liệu. Đồng thời, SAS hay SPSS giúp xử lý tốt công việc. Ngoài ra, Power Bi, Tableau, Rapidminer, Matlab, Excel, Gephi cũng là những công cụ đắc lực của một DE.

## Công việc chính của một Data Engineer

### Crawling – Thu thập và lưu trữ dữ liệu

Để thực hiện công việc này, các DE sẽ phối hợp cùng với DBA (Database Administration) để thiết kế các vùng lưu trữ dữ liệu từ các nguồn sao cho hiệu quả nhất. Đồng thời, sau khi DBA thực hiện quản lý, sao lưu và phục hồi dữ liệu từ các database, file server, DE sẽ là người đưa dữ liệu vào các Database và file server. Công việc này có thể được thực hiện thông qua FTP, drag, drop… và chính DE sẽ là người quyết định lữu trữ những dữ liệu này ra sao, lưu trữ dưới định dạng nào như .csv, .xlsx, json, .dat, database…

Web Crawling là quá trình thu thập thông tin từ các Website trên mạng Internet theo các đường links cho trước. Các Web Crawler sẽ truy cập các links này để download toàn bộ nội dung của trang web cũng như tìm kiếm thêm các đường link bên trong để tiếp tục truy cập và download nội dung từ các đường link này. Dữ liệu sau khi được tải về sẽ được đánh chỉ số (indexing) rồi lưu vào cơ sở dữ liệu. Ví dụ như đối với trang amazon.com, Web Crawling sẽ thu thập toàn bộ nội dung của trang web này (tên các sản phẩm, thông tin chi tiết, bảng giá, hướng dẫn sử dụng, các reviews và comments về sản phẩm…).

Các công cụ phổ biến được DE sử dụng trong quá trình thu thập và lưu trữ dữ liệu bao gồm: Bash jobs, SQL, ETL Tools…

### ETL – Lưu chuyển, làm sạch và chuẩn hóa dữ liệu

ETL là viết tắt của Extract – Transform – Load, tạm được dịch là Trích xuất – biến đổi – tải. Đây là quy trình chung dùng để sao chép dữ liệu từ một hay nhiều nguồn vào hệ thống đích đại diện cho dữ liệu khác với nguồn dữ liệu. Trong quy trình ETL, dữ liệu sẽ được trải qua các bước:

* Extract/ Trích xuất: Đây là quá trình đọc dữ liệu từ cơ sở dữ liệu. Trong giai đoạn này, nó chọn dữ liệu nguồn hoặc siêu dữ liệu và dữ liệu kinh doanh đến từ các nguồn khác nhau. Dữ liệu có thể ở bất kỳ định dạng nào như tệp xml, tệp phẳng hoặc bất kỳ tệp cơ sở dữ liệu nào. Tại giai đoạn này, dữ liệu gốc được kiểm tra. Dữ liệu sẽ được kiểm tra để xem có bất kỳ lỗi nào hay không. Giai đoạn Extarct sẽ kiểm tra các lỗi hoặc sự không nhất quán của dữ liệu bằng cách sử dụng một số kỹ thuật trí tuệ nhân tạo. Trong ngắn hạn, nó xác minh xem chất lượng của sản phẩm có đạt hay không. Sau đó dữ liệu sẽ được chuyển tiếp tới giai đoạn tiếp theo để tiếp tục được xử lý.
* Transform/ Biến đổi: Đây là giai đoạn thứ hai trong ETL. Chuyển đổi là quá trình trong đó định dạng ban đầu được chuyển đổi thành định dạng bắt buộc mà người dùng muốn. Chuyển đổi là mô hình hóa hoặc thay đổi dữ liệu theo yêu cầu của người dùng. Ví dụ những thay đổi có thể là thay đổi về số cột hoặc số hàng.
* Load/ Tải: Tải dữ liệu và xác nhận số lượng hàng đã được xử lý. Quá trình này còn được biết tới như quá trình ghi chép dữ liệu vào cơ sở dữ liệu đích.

Như vậy, với những DE chuyên thực hiện công việc ETL, bạn sẽ tiến hành lưu chuyển, trao đổi dữ liệu dữ các nơi lưu trữ khác nhau nhằm mục đích đối chiều, bổ sung, thậm chí là backup dữ liệu. Đồng thời, các DE thực hiện làm sạch và chuẩn hóa dữ liệu bằng cách loại bỏ dữ liệu dư thừa, dữ liệu rác và chuyển các dữ liệu về các định dạng nhất định. Sau khi làm sạch, dữ liệu được đưa về tập trung, load lên một kho chứa duy nhất là Data Warehouse. Khi dữ liệu đã được lưu trữ tại đây có nghĩa là dữ liệu đã được thiết kế theo mô hình chuyên biệt sẵn sàng cho việc phân tích hoặc khôi phục dữ liệu trong trường hợp cần thiết.

Các công cụ giúp các DE có thể thực hiện quá trình ETL có thể là: SQL, SSIS, Data Stage, Bash jobs, Python, Scala, DataBricks…

### Visualazation – Phân tích và trực quan hóa dữ liệu

Một công việc không thể thiếu của DE đó là phân tích và trực quan hóa dữ liệu. Đối với các công việc về phân tích dữ liệu nhiều người thường nghĩa ngay tới nghề DA (Data Analyst), tuy nhiên đối với DE, việc biết về phân tích và trực quan hóa dữ liệu là điều hết sức cần thiết. Từ những dữ liệu được lưu trữ tại các Data Warehouse hay các Database, các DE sẽ sử dụng các công cụ như BI Tools để lấy dữ liệu có ý nghĩa và tiến hành trực quan hóa thông qua các dashboard và phân tích. Đối với những yêu cầu truy vấn đưa ra phân tích phức tạp hơn, hay những dữ liệu khó trực quan hóa hơn, DE cần áp dụng các mô hình thuật toán, sử dụng các code độc lập hoặc nhúng dữ liệu vào các tool để thực hiện trực quan hóa dữ liệu từ đó đưa ra phân tích và đánh giá dữ liệu.

Các công cụ phổ biến được các DE sử dụng để thực hiện Visualization bao gồm: R, Python, Oracle BI, Power BI, Tableau…

# Tổng quan

## Cào dữ liệu

### Mô tả công việc cào dữ liệu

Crawl là cào dữ liệu (Crawl Data) là một thuật ngữ không còn mới trong Marketing, và SEO. Vì Crawl là kỹ thuật mà các con robots của các công cụ tìm kiếm sử dụng như: Google, Bing Yahoo…

Công việc chính của crawl là thu thập dữ liệu từ một trang bất kỳ. Rồi tiến hành phân tích mã nguồn HTML để đọc dữ liệu. Và lọc ra theo yêu cầu người dùng hoặc dữ liệu mà Search Engine yêu cầu.

Trước tiên, thu thập dữ liệu webpage tại các URL. Từ các page này, tìm thấy các siêu liên kết đến nhiều URL khác và thêm các liên kết mới tìm được vào danh sách các trang cần thu thập thông tin tiếp theo. Sau đó sẽ truy cập vào URL các trang có thông tin quan trọng và bắt đầu cào những dữ liệu quan trọng phù hợp với yêu cầu.

### Phần mềm sử dụng

Phần mềm: Visual Code

Ngôn ngữ lập trình: Python

Thư viện hỗ trợ: Beautifulsoup, Pandas

## Mục 2.2

# TRIỂN KHAI

## Cào dữ liệu

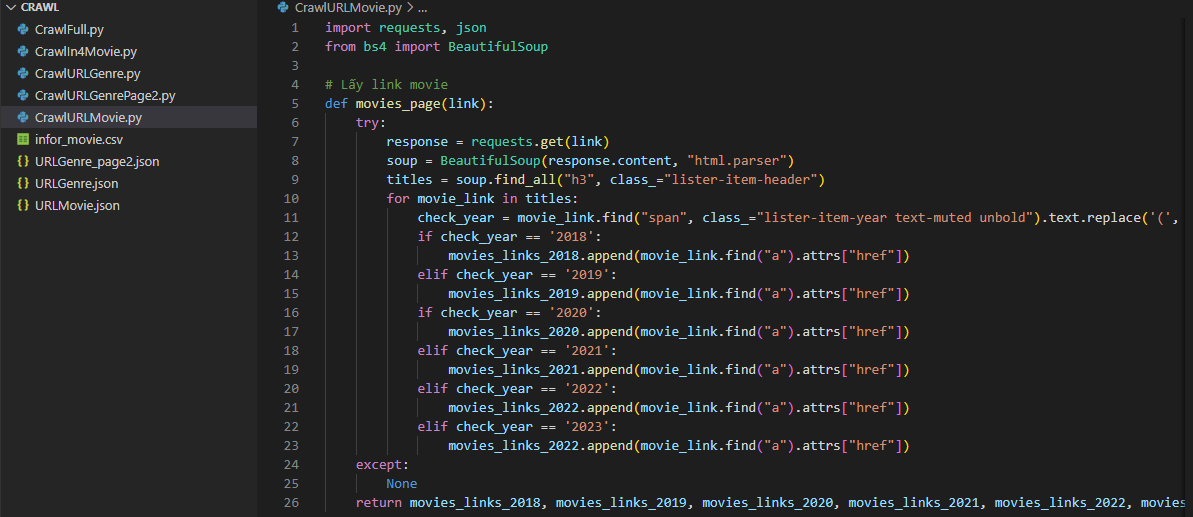
* Bước 1: Crawling URL từng thể loại thể loại phim (trang 1) và đưa vào file json (URLGenre.json)

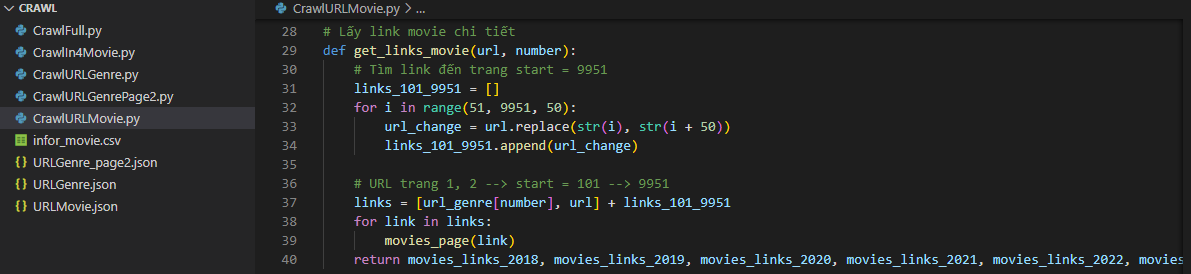


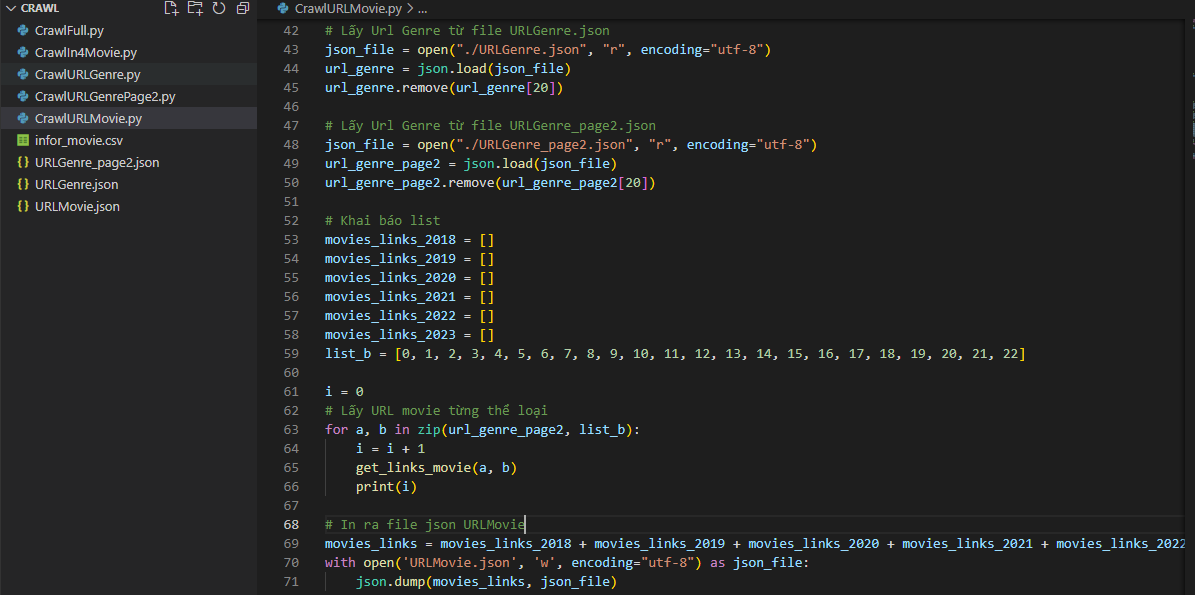
* Bước 2: Từ file json (URLGenre.json) lấy URL đưa vào và tiếp tục lấy URL thể loại phim (trang 2) và đưa vào file json URLGenre\_page2.json



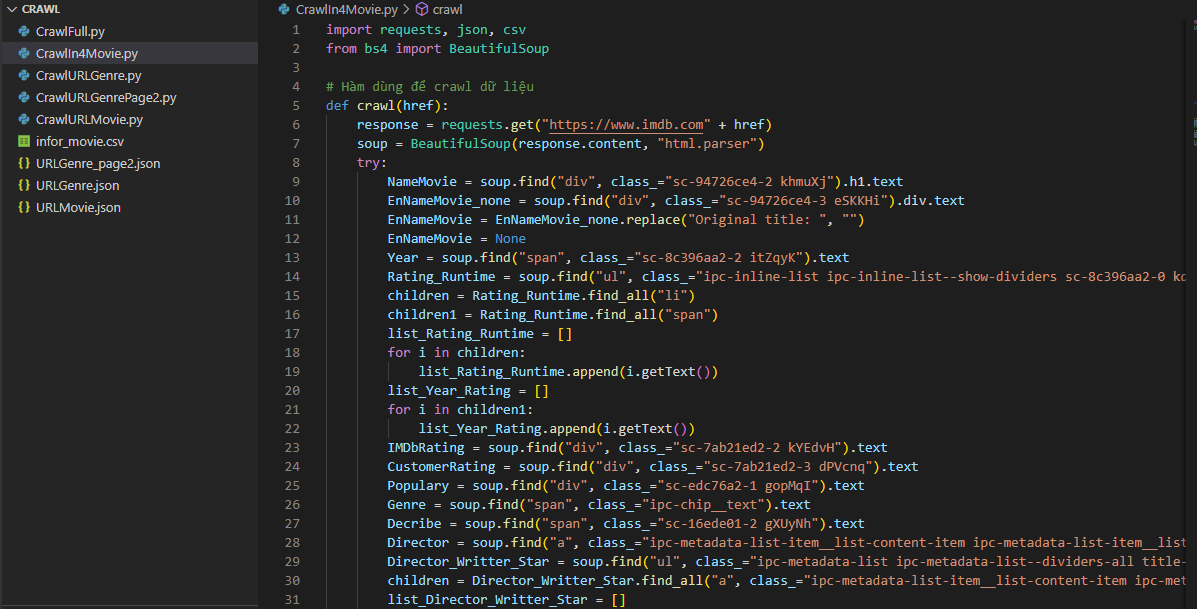
* Bước 3: Từ file json (URLGenre.json) lấy URL từng bộ phim dẫn đến trang phim và đưa vào URLMovie.json

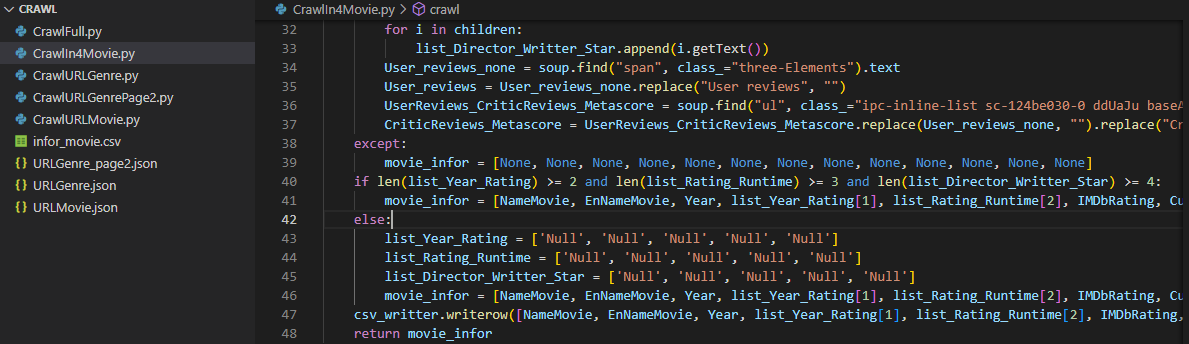


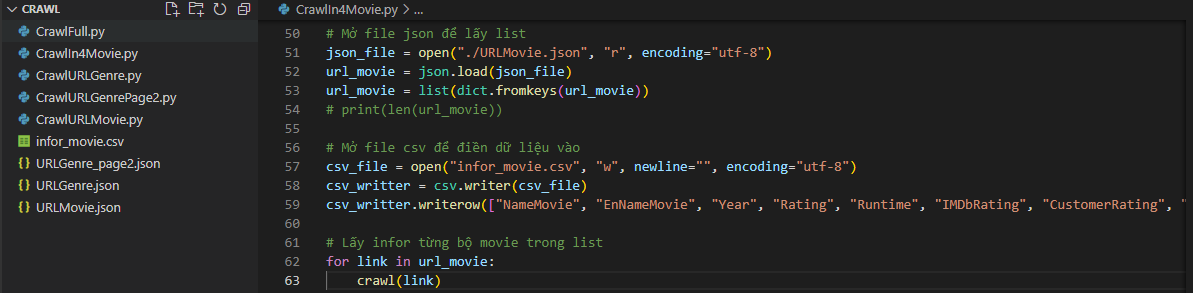




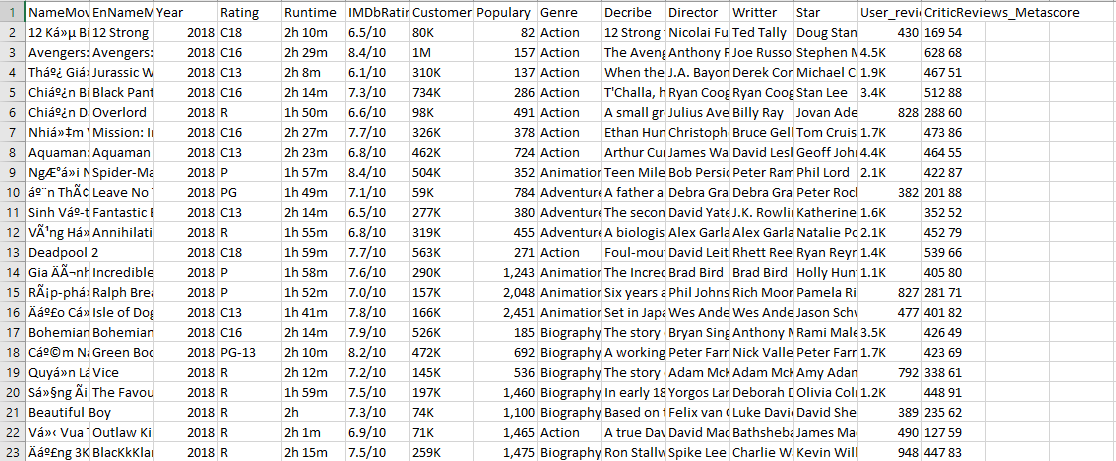
* Bước 4: Từ file json (URLMovie.json) lấy ra URL và truy cập vào trang phim để lấy thông tin sau đó xuất ra file CSV (infor\_movie.csv)



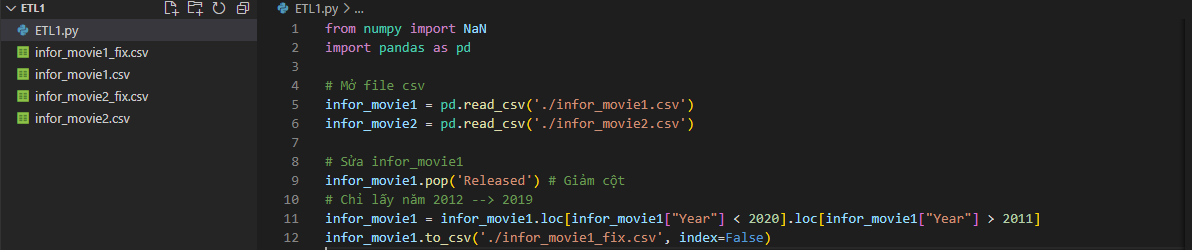


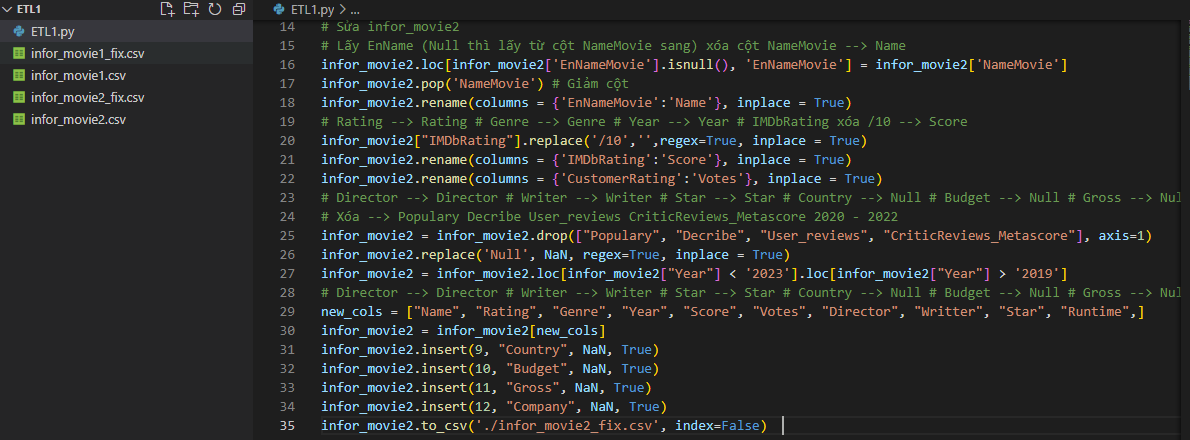


* File CSV sau khi crawling

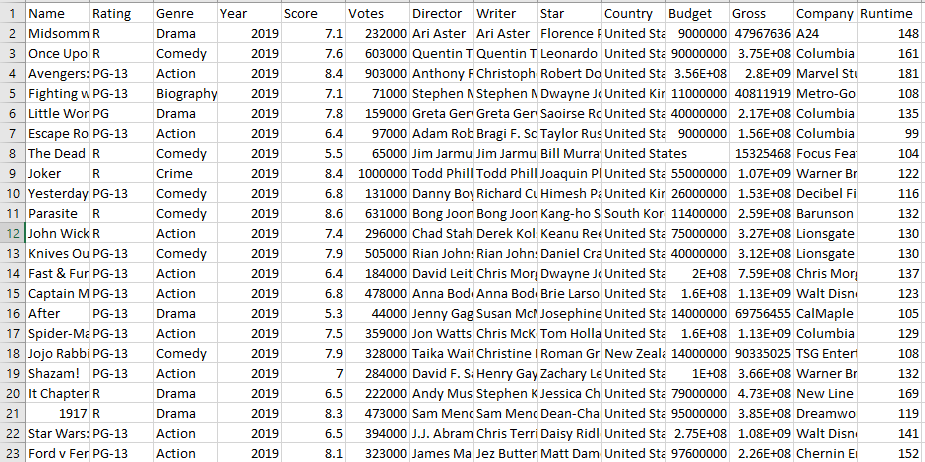


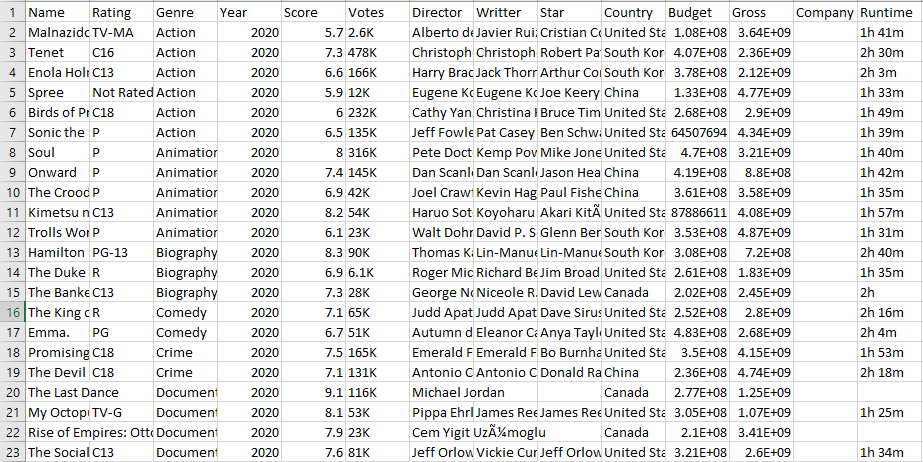
* Bước 5: Phân tách các cột và trường để thuận tiện cho việc xử lí dữ liệu





* Bảng sau khi phân tách hoàn chỉnh, hoàn thành công việc crawl dữ liệu và thu được 2 bảng dữ liệu (infor\_movie1\_fix.csv, infor\_movie2\_fix.csv) sau đó đưa đến cho ETL bắt đầu công việc xử lý





## Mục 3.2

# KẾT QUẢ

## Mục 4.1…

## Mục 4.2…

Kết quả được xây dựng dựa trên ngôn ngữ lập trình thể hiện như (Bảng 4.1).

Bảng . Kiến trúc

|  |  |  |
| --- | --- | --- |
| Lớp | Ý nghĩa | Tham số |
|  |  |  |
|  |  |  |
|  | | |

Hình . Kiến trúc của mô hình

# KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Đề tài đã thực hiện được …

# TÀI LIỆU THAM KHẢO

* + - 1. Tên tác giả, Tên tác giả… *Tên công trình*. Nơi xuất bản, năm, trang
      2. Tên tác giả, Tên tác giả… *Tên công trình*. Nơi xuất bản, năm, trang

# PHỤ LỤC