**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**

**XỬ LÝ VÀ PHÂN TÍCH XU HƯỚNG QUẢNG CÁO GAME QUA VIDEO**

Sinh viên thực hiện : Trương Thị Ngọc Viên

Nguyễn Bá Khánh

Lớp : 45K21.1

Đơn vị thực tập : TMA Solution

Cán bộ hướng dẫn : Nguyễn Thanh Bình

Lê Bá Thiện

Giảng viên hướng dẫn : Ths.Cao Thị Nhâm

**Đà Nẵng, 8/2022**

NHẬN XÉT CỦA ĐƠN VỊ THỰC TẬP

Họ và tên sinh viên:

Lớp: Khoa: Trường:

Thực tập từ ngày: …./……/ 2022 đến ngày: ........./ .……./ 2022

Tại:

Địa chỉ:

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:

……….., ngày .......tháng ......năm 2022

Xác nhận của đơn vị thực tập

# LỜI CẢM ƠN

# LỜI CAM ĐOAN

**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**

[LỜI CẢM ƠN i](#_Toc108821186)

[LỜI CAM ĐOAN ii](#_Toc108821187)

[DANH MỤC HÌNH ẢNH vi](#_Toc108821188)

[DANH MỤC BẢNG BIỂU vii](#_Toc108821189)

[DANH MỤC CÁC TỪ VIẾT TẮT viii](#_Toc108821190)

[LỜI MỞ ĐẦU 1](#_Toc108821191)

[CHƯƠNG 1. TỔNG QUAN LÝ THUYẾT 2](#_Toc108821192)

[**1.1** **Azure và Azure DataBricks** 2](#_Toc108821193)

[1.1.1 Microsoft Azure 2](#_Toc108821194)

[1.1.2. Databrick sử dụng nền tảng Azure 7](#_Toc108821195)

[***1.2.*** ***Data Egineer*** 7](#_Toc108821196)

[1.2.1. Giới thiệu tổng quan về data egineer 7](#_Toc108821197)

[1.2.2. Thu thập dữ liệu - Crawling 7](#_Toc108821198)

[1.2.3. Xử lý dữ liệu – ETL (Extract - Transform – Load) 7](#_Toc108821199)

[1.2.4. Trực quan hoá và phân tích dữ liệu bằng Power BI 9](#_Toc108821200)

[CHƯƠNG 2. TRIỂN KHAI … 11](#_Toc108821201)

[2.1. Mục 3.1 11](#_Toc108821202)

[2.1.1. Mục 3.1.1 11](#_Toc108821203)

[2.1.2. Mục 3.1.2 11](#_Toc108821204)

[2.2. Mục 3.2 11](#_Toc108821205)

[CHƯƠNG 3. KẾT QUẢ 12](#_Toc108821206)

[3.1. Mục 4.1… 12](#_Toc108821207)

[3.2. Mục 4.2… 12](#_Toc108821208)

[KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 13](#_Toc108821209)

[TÀI LIỆU THAM KHẢO 14](#_Toc108821210)

[PHỤ LỤC 15](#_Toc108821211)

# 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**

# TỔNG QUAN LÝ THUYẾT

* 1. **Azure và Azure DataBricks**
     1. ***Microsoft Azure***

1. **Tổng quan về Azure và nền tảng điện toán đám mây**

Hiện nay các nền tảng điện toán đám mây đang phát triển nhanh chóng và đóng vài trò quan trọng trong cơ sở hạ tầng công nghệ thông tin. Hai cái tên luôn được nhắc đến đầu tiên là Amazon Web Services và Microsoft Azure, tuy không được phủ sóng rộng như Amazon Web Services nhưng Azure vẫn được đánh giá cao với tốc độ phát triển vượt bật

Microsoft Azure là một nền tảng điện toán đám mây công cộng, cung cấp giải pháp gồm Cơ sở hạ tầng dưới dạng dịch vụ - (IaaS), nền tảng dưới dạng dịch vụ (PaaS) và phần mềm dưới dạng dịch vụ (SaaS). Azure có thể được sử dụng cho các dịch vụ như phân tích, ‘ảo hóa’ máy tính, lưu trữ mạng, …

* **Cơ sở hạ tầng dưới dạng dịch vụ - Infrastructure as a Service (IaaS)**: Azure cho người sử dụng xây dựng và tự quản lý các máy ảo, không cần quan tâm đến viêc quản lý server. Giảm bớt chi phí vốn để tập trung hơn vào vấn đề kinh doanh.
* **Nền tảng dưới dạng dịch vụ - Platform as a Service (PaaS)**: Azure quản lý hết từ các máy ảo được người dùng xây dựng đến Server, tự động thiết đặt thêm các dịch vụ cho người dùng. Chủ yếu dùng để lưu trữ database với tính năng backup cùng tính hữu dụng cao.
* **Phần mềm dưới dạng dịch vụ - Software as a Service (SaaS):** Cung cấp dịch vụ thông qua các phần mềm, người dùng tải phần mềm để sử dụng các dịch vụ. Các hệ thống đi kèm được lưu trữ trực tuyến và được quản lý toàn bộ bởi Azure.

Microsoft Azure có tốc độ hoạt đông rất nhanh, độ linh hoạt cao và giá cả phải chăng. Điều đó giúp cho Azure trở thành một dịch vụ đám mây công cộng tốt nhất trên thi trường hiện nay với một số dặc điểm chính:

* **Linh hoạt**: Di chuyển (tải về hoặc tải lên) tài nguyên máy tính qua lại máy chủ khi hoạt độn
* **Mở**: Hỗ trợ hầu hết mọi hệ điều hành, ngôn ngữ, công cụ
* **Đáng tin cậy**: Thỏa thuận mức dịch vụ (SLA) là 99,95% với độ tin cây cao cùng khả năng quản lý dịch vụ và bảo mật thông tin tốt.
* **Toàn cầu**: Dữ liệu được đặt trong các trung tâm dữ liệu mà người dùng có thể truy cập từ bất cứ nơi nào chỉ bằng Internet.
* **Tiết kiệm**: Chỉ trả tiền cho những gì bản sử dụng, giảm đi hoặc không còn các chi phí triển khai ban đầu.

Với nền tảng điện toán đám mây việc xây dựng các tài nguyên trên đó trở nên dễ dàng hơn:

* Ít chi phí hoặc không có chi phí ban đầu
* Không cẩn phải quản lý các khía cạnh như: Máy chủ vật lý, mạng, hệ thống làm lạnh, hệ thống phòng vệ, … vì đã có Azure hỗ trợ
* Chi trả khi sử dụng, hoặc tính phí khi chuyển sang các dịch vụ khác.

Những nền tảng điện toán đám mây mà Azure cung cấp:

* Public Cloud: Cung cấp các dịch vụ sử dụng thông qua web browser, chi phí thấp hơn, ít bảo dưỡng, các dịch vụ được cung cấp và quản lý bởi Microsoft Azure
* Private Cloud: Các cơ sở hạ tầng thuộc quyền quản lý của công ty, chính phủ hoặc các tổ chức tài chính, khi họ muốn có tính sở hữu cao, và cá nhân đối với các tài nguyên mà họ xây dựng trên đó, vì vậy tốn nhiều chi phí hơn.
* Hybrid Cloud: Kết nối Azure đến cơ sở hạ tầng trên máy của người dùng. Thừa hưởng cả hai điểm lợi từ việc xây dựng on premise và cloud, có tính linh hoạt cao. Nhưng cần thêm chi phí và bảo trì cho các tài nguyên được xây dựng

1. **Xây dựng tài nguyên trên Microsoft Azure**
2. **Resources và Resource group**

Với sự hổ trợ của Azure việc xây dựng các tài nguyên thông qua xây dựng máy ảo và quản lý nó trở nên đơn giản về dễ dàng hơn nhiều với người dùng hiện nay. Việc tạo ra các tài nguyên trong Azure tức ra tạo ra các Data center (Container) để lưu trữ các cơ sở hạ tầng được người dùng thiết đặt về sau. Các Data center liên kết với nhau với độ trễ thấp và tạo thành một vùng (Region)

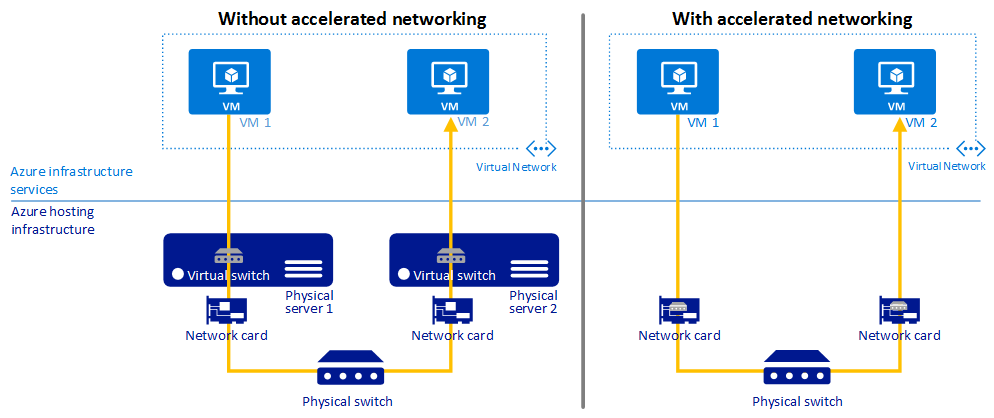
Các tài nguyên được xây dựng có thể là Website, máy ảo, cơ sở dữ liệu. Các tài nguyên được xây dựng có thể liên kết đến các tài nguyên khác tạo nên mạng lưới các tài nguyên. Để dễ dàng trong việc quản lý nhiều tài nguyên người dùng có thể gom các tài nguyên thành nhóm (Resource group).

Các máy ảo khi xây dựng được lưu trữ trong mạng ảo (Virtual Network). Mạng ảo bao gồm nhiều mạng con (Subnet). Máy ảo khi tạo nên sẽ được liên kết đến một giao diện mạng (Interface Network) có nhiệm vụ chính là cung cấp các địa chỉ Ip để các máy ảo có thể liên lạc được với nhau. Mỗi mạng ảo được chỉ định một địa chỉ IP cụ thể, các mạng con được cung cấp Ip thông qua Ip của mạng ảo.

Có 2 loại địa chỉ Ip được phân bổ cho các máy ảo:

* **Private Ip**: Ip được cung cấp thông qua Ip của các mạng con, thông qua Private Ip các máy ảo trong cùng 1 mạng có thể giao tiếp được với nhau
* **Public Ip:** phân bổ cho các máy ảo để giao tiếp với các máy ảo khác qua Internet

Trên thực tế việc giao tiếp giữa các mạng là không thể, phải tạo ra các kết nối mạng có thể nhìn thấy được (Virtual network Peering connection) từ đó các máy ảo có thể liên kết được với nhau



Các máy ảo được xây dựng có thể gắn thêm các ổ đĩa (Disk) mặc định là các ổ đĩa OS (VHD files) các ổ đĩa này có thể được gắn thêm vào. Ngoài chức năng lưu trữ, các ổ đĩa này còn có thể:

* Dừng máy ảo bất cứ khi nào mà người dùng cần
* Kiểm tra lưu lượng sử dụng mạng
* Xem các số liệu cơ bản về CPU, mạng

1. **Quản lý máy ảo**

Mạng máy ảo khi vào hoạt động sẽ được giám sát và quản lý thông qua một dịch vụ của Microsoft Azure là Network Security Groups (NSG). Các chức năng chính của NSG:

* Giúp người dùng có thể giám sát lưu lượng vào các mạng con hoặc các máy ảo
* Thiết lập luật bên trong và bên ngoài lưu thông
* Cho phép những lưu thông nào rõ ràng (đến từ cổng nào, giao thức kết nối nào, thực hiện hành động nào?) mới có thể đi qua và kết nối với các máy ảo
* Những luật được xây dựng phải có nguồn, đích

Các ứng dụng (ví dụ: Website, app, …) có thể được xây dựng trên một máy ảo nào đó. Khi người dùng muốn truy cập vào Website cũng đồng nghĩa truy cập đến máy ảo. Đối với một cá nhân thì không ảnh hưởng, nhưng khi nhiều cá nhân cùng truy cập vào thì việc xử lý tải sẽ trở nên khó khăn. Bộ mở rộng quy mô (Scale Sets) là một giải quyết cho vấn đề này. Các máy ảo khi đó sẽ được quản lý theo quy mô của bộ dịch vụ, bộ quy mô sẽ tự động thêm máy ảo để mở rộng quy mô phục vụ cho nhu cầu truy cập vào của người dùng.

Bộ mở rộng (Scale Sets) thực hiện các chức năng như:

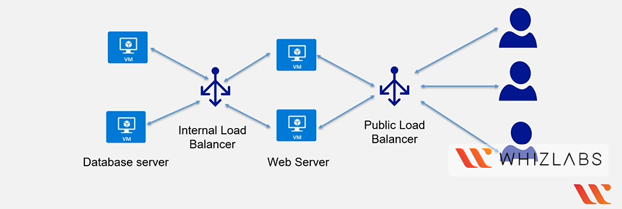
* Là dịch vụ cho phép tạo ra và thiết đặt những máy ảo giống hệt nhau
* Cho phép mở rộng quy mô theo yêu cầu
* Nâng cao tính sẵn có cho cấu trúc cơ sở hạ tầng được xây dựng
* Có thể đặt máy ảo sau bộ mở rộng hoặc bộ cân bằng tải

Một dịch vụ quản lý khác cũng được Azure cung cấp cho người dùng giúp quản lý lưu lượng, chuyển yêu cầu của người dùng phân bổ đều cho các máy ảo phụ trợ tránh trường hợp một máy ảo phụ trợ bị quá tải yêu cầu truy cập thông qua bộ cân bẳng tải (Azure Load Balancer).

Có 2 loại cân bằng tải người dùng có thể tạo ra để quản lý lưu lượng trên Azure bao gồm:

* Internal Load Balancer: cân bằng tải nội bộ điều phối lưu lượng giữa máy ảo với nhau
* Public Load Balancer: cân bằng tải giữa người dùng với máy ảo

Ngoài ra còn có những nhóm phụ trợ khác (đây cũng là những máy ảo với chức năng khác) giúp xác định máy ảo, thiết đặt bộ mở rộng, xác định những giao thức, số cổng và khoảng thời gian cho mỗi lần tải.



1. **Azure Database**

Azure SQL Database là nên tảng dịch vụ được quản lý bởi Azure được triển khai linh hoạt với nhiều mức giá, giúp truy cập đến các chức năng của Microsoft SQL Server. Khi triển khai SQL trên Virtual Machine sẽ có đầy đủ các chức năng, dễ dàng vận chuyển data từ máy của người dùng lên nền tảng đám mây. Sử dụng các công cụ truyền thống như lưu trên cơ sở dữ liệu dưới máy của người dừng và có thể kết nối bằng SQL Server Management

Azure SQL Data Warehouse: dùng để lưu trữ dữ liệu lớn. Dữ liệu trong data Warehouse được lưu trữ trong các bảng quan hệ và dùng cách lưu theo tên cột. Sử dụng dịch vụ lưu trữ ở Warehouse giúp giảm bớt các chi phí, cải thiện hiệu suất truy vấn. Tại đây lưu trữ các dữ liệu sạch sử dụng cho mục đích phân tích và báo cáo.

Một dịch vụ khác được Azure cung cấp là dịch vụ không máy chủ (Serverless). Đối với giải pháp lưu trữ này người dùng không cần quan tâm đến máy chủ, cũng như vấn đề bảo trì (một ví dụ điển hình là DataBricks), dữ liệu người dùng được đưa lên từ máy tính cá nhân nhưng thật sự không được lưu trữ trên server.

### Databrick sử dụng nền tảng Azure

Với sự phát triển xã hội hiện nay, đặc biệt là trong xu hướng chuyển đổi số của cuộc cách mạng công nghiệp 4.0, mọi hoạt động dần dần được được internet hóa. Lượng thông tin trao đổi trên Internet ngày càng đa dạng và không ngừng tăng. Chính vì vậy, dữ liệu ngày càng lớn và đa dạng. Big Data lớn về số lượng, dữ liệu phức tạp có thể có cấu trúc hoặc không có cấu trúc (noSQL). Những yếu tố này làm cho Big Data khó bắt giữ lại, khai phá và quản lý nếu dùng các phương thức truyền thống.

Dù mới phát triển gần đây nhưng Databricks là một nền tảng khoa học dữ liệu, kỹ thuật dữ liệu và phân tích dữ liệu cộng tác dựa trên đám mây, kết hợp tốt data warehouses,  data lakes và lakehouse với nhau.

Databrick là nơi hỗ trợ xử lý big data được xây dựng từ Apache Spark hỗ trợ xử lý dữ liệu với tốc độ nhanh và toàn vẹn. Databrick hỗ trợ trên ba nền tảng lớn hiện nay là Microsoft Azure, amazon web services và google cloud. Azure Databricks là dịch vụ triển khai Databricks trên nền tảng Azure, cung cấp khả năng autoscale, tương tác với các thành viên khác dễ dàng thông qua workspace. Azure Databricks hỗ trợ nhiều ngôn ngữ như Java, Python, Scala, R,... Tuy nhiên, vì được xây dựng dựa trên lõi là Apache Spark nên Databricks hỗ trợ mạnh mẽ cho ngôn ngữ Scala với tốc độ xử lý nhanh và kho thư viện lớn.

Azure Databricks hiện đã trở thành một nền tảng phân tích dữ liệu được tối ưu hóa cho nền tảng dịch vụ đám mây Microsoft Azure. Azure Databricks cung cấp ba môi trường:

* Databricks SQL
* Databricks data science and engineering
* Databricks machine learning

Để sử dụng databrick, trước tiên cần tạo một cluster, sau đó đẩy file data vào DBFS để tạo môi trường làm việc. Sau khi chuẩn bị xong tạo một notebook mới để tiến hành xử lý dữ liệu.

## Data Egineer

### Giới thiệu tổng quan về data egineer

Ngày nay với sự phát triển không ngừng của khoa học, công nghê. Khi xã hội ngày càng phát triển thì càng có những công việc xuất hiện nhiều hơn. Đặc biệt là công việc liên quan đến công nghệ thông tin. Data engineer – kỹ sư dữ liệu là một thuật ngữ để chỉ những người làm việc trong lĩnh vực dữ liệu. Kỹ sư dữ liệu là người xây dựng một systems tông hợp lưu trữ và xuất dữ liệu từ một ứng dụng, hệ thống đã được tạo ra trước đó. Các kỹ sư dữ liệu thực hiện các công việc chính:

* Làm việc liên quan đến hệ thống điện toán đám mây
* Lập trình
* Trao dồi kiến thức về các công cụ mới
* Thực hiện các công việc tạo ETL/data piplines
* Cấu trúc dữ liệu nâng cao

Những công cụ mà kỹ sư dữ liệu thường sử dụng là R và Python. Họ sẽ phải thường xuyên sử dụng các packages như ggplot2 khi làm việc với các công cụ này, trực quan hóa dự liệu với R hoăc thao tác với thư viện Pandas Python. Một phần khác các công cụ như SAS hay SPSS sẽ xử lý tốt các công việc liên quan đến dữ liệu

Data engineer thực hiện thu thập và lưu trữ dữ liệu được thiết kế và thực hiện, xử lý trước dữ liệu và chuyển đổi nó thành một định dạng có thể sử dụng được. Họ xây dựng dữ liệu và đảm bảo rằng dữ liệu trôi chảy. Để thực hiện được những điều này Data engineer cần có những kỹ năng cơ bản sau:

* Kỹ năng lập trình cơ bản
* Khả năng tập trung cao độ
* Kỹ năng phân tích logic
* Kỹ nẵng thiết kế báo cáo
* Kỹ năng giao tiếp, trình bày
* Biết cách quản lý thời gian, sắp xếp công việc
* Có thể kể câu chuyện từ số liệu

### Thu thập dữ liệu – Crawling

Thu thập dữ liệu (Web Scraping) là thuật ngữ chỉ việc tìm kiếm và thu thập các thông tin trên một website. Quét hay cào dữ liệu được sử dụng khá phổ biến trong hoạt động SEO web thuộc lĩnh vực marketing online hay trong lĩnh vực công nghệ thông tin.

Kỹ thuật Crawl dữ liệu có ý nghĩa quan trọng đối với website. Nó thực hiện việc thu thập dữ liệu từ một website nào đó rồi tổng hợp, phân tích và bóc tách thông tin theo yêu cầu mà người dùng đặt ra hay theo các dữ liệu mà công cụ tìm kiếm yêu cầu. Sau khi thực hiện xong, dữ liệu sẽ được lưu vào database.

Crawl giúp lấy thông tin từ website qua đường link web được nhập vào, thông qua các lệnh được lập trình bởi một ngôn ngữ lập trình nào đó (Python, Java, C#) để lấy dữ liệu. Việc crawl dữ liệu cũng tương tự như hoạt động của robots của các công cụ tìm kiếm (Google, Bing, Baidu,…) thường làm để phục vụ cho việc tìm kiếm.

Những ưu điểm của việc thực hiện crawl data:

* Giảm thiểu thời gian và công sức cho quá trình thu thập khối lượng lớn thông tin và dữu liệu
* Có thể dễ dàng tổng hợp thông tin từ nhiều trang web khác nhau. Sau đó thực hiện đối chiếu, so sánh và làm các thuật toán để phân tích thị trường, tiết kiệm thời gian, công sức và hiệu quả hơn.
* Thông qua crawl giúp lấy thông tin nhanh chóng, chính xác và hữu hiệu

Dù khá hiệu quả trong việc lấy dữu liệu nhưng Crawl vẫn có một số các hạn chế:

* Khi website thay đổi cấu trúc HTML thì phải thực hiện update, cập nhật để chương trình crawl phù hơp với những thay đổi đó.
* Trong thực tế việc crawl có thể coi là vi phạm các quyền về bảo mật thông tin, nên các trang web uy tín và có chất lượng thường sẽ không thực hiện crawl được

### Xử lý dữ liệu – ETL (Extract - Transform – Load)

ETL là viết tắt của trích xuất, biến đổi và tải, là một quá trình tích hợp dữ liệu kết hợp dữ liệu từ nhiều nguồn dữ liệu vào một kho dữ liệu nhất quán, duy nhất được tải vào  kho dữ liệu  hoặc hệ thống đích khác. ETL làm sạch và sắp xếp dữ liệu theo cách giải quyết các nhu cầu thông minh kinh doanh cụ thể, như báo cáo hàng tháng, nhưng nó cũng có thể giải quyết các phân tích nâng cao hơn, có thể cải thiện quy trình back-end hoặc trải nghiệm người dùng cuối. ETL thường được một tổ chức sử dụng để:

* Trích xuất dữ liệu từ các hệ thống kế thừa
* Xóa dữ liệu để cải thiện chất lượng dữ liệu và thiết lập tính nhất quán
* Tải dữ liệu vào cơ sở dữ liệu đích

Dữ liệu thô sau khi crawl cần được ETL để trở thành một dữ liệu sạch, toàn vẹn và có ý nghĩa. Dữ liệu thô thường sẽ chứa các giá trị Null, duplicate, negative,... Tuỳ vào mỗi dataset mà cần tiến hành xử lý dữ liệu phù hợp. Công cụ để tiến hành ETL cũng khá đa dạng và phụ thuộc vào yêu cầu cũng như mục đích sử dụng của người dùng cuối. Tuy nhiên, những công cụ phổ biến được dùng để ETL hiện nay là ngôn ngữ: R, Python, Scala, Java, SQL,...

Trong thời đại 4.0 lượng dữ liệu đồ sộ khiến cho việc xử lý làm sạch phải diễn ra trên môi trường cloud tích hợp đa ngôn ngữ với bộ nhớ lớn đảm bảo quá trình xử lý dữ liệu diễn ra nhanh, tiện, chính xác. Hiện nay, có rất nhiều nền tảng cloud phục vụ cho việc ETL big data, nhưng lớn nhất phải kể đến AWS, Azure và Google Cloud.

Dữ liệu thô sau khi được trích xuất sẽ trải qua quá trình xử lý dữ liệu. Tại đây, dữ liệu được chuyển đổi và hợp nhất cho trường hợp sử dụng phân tích dự kiến ​​của nó. Giai đoạn này có thể bao gồm các nhiệm vụ sau:

* Lọc, làm sạch, bỏ trùng lặp, xác thực dữ liệu.
* Thực hiện các phép tính, dịch hoặc tóm tắt dữ liệu dựa trên dữ liệu thô như thay đổi tiêu đề hàng và cột để có tính nhất quán, chuyển đổi kiểu dữ liệu, chỉnh sửa chuỗi văn bản, v.v.
* Thực hiện kiểm tra để đảm bảo chất lượng và toàn vẹn của dữ liệu.
* Xóa, mã hóa hoặc bảo vệ dữ liệu do các cơ quan quản lý ngành hoặc chính phủ quản lý.
* Định dạng dữ liệu thành các bảng hoặc bảng nối để phù hợp với lược đồ của data warehouse.

Dữ liệu tại data warehouse thuộc lớp gold có thể đã có ý nghĩa có thể phục vụ cho mục đích cao hơn của data science và data analyst. Nếu quá trình ETL không tốt,sẽ ảnh hưởng lớn đến việc nghiên cứu, phân tích và đánh giá dữ liệu.

### Phân tích và trực quan hoá dữ liệu bằng công cụ Power BI

Trong thời đại nhiều dữ liệu của chúng ta, việc hiểu cách phân tích và rút ra ý nghĩa thực sự từ những hiểu biết kỹ thuật số của doanh nghiệp chúng ta là một trong những động lực chính dẫn đến thành công.

Phân tích dữ liệu là quá trình thu thập, mô hình hóa và phân tích dữ liệu để rút ra những hiểu biết sâu sắc hỗ trợ việc ra quyết định. Có một số phương pháp và kỹ thuật để thực hiện phân tích tùy thuộc vào ngành và mục đích của cuộc điều tra. Tất cả các phương pháp khác nhau này chủ yếu dựa trên hai lĩnh vực cốt lõi: nghiên cứu định lượng và định tính.

Với sự trợ giúp của các kỹ thuật khác nhau như phân tích thống kê, hồi quy, mạng nơ-ron, phân tích văn bản, v.v., bạn có thể bắt đầu phân tích và thao tác dữ liệu của mình để rút ra các kết luận có liên quan. Ở giai đoạn này, bạn tìm thấy các xu hướng, mối tương quan, các biến thể và các mẫu có thể giúp bạn trả lời các câu hỏi bạn nghĩ đến đầu tiên trong giai đoạn xác định. Các công nghệ khác nhau trên thị trường hỗ trợ các nhà nghiên cứu và người dùng doanh nghiệp trung bình quản lý dữ liệu của họ. Một số trong số đó bao gồm phần mềm trực quan và trí tuệ kinh doanh, phân tích dự đoán, khai thác dữ liệu, trong số những phần mềm khác.

Phân tích dữ liệu có 4 mức độ chính từ thấp đến cao gồm:

* Phân tích mô tả: cho chúng ta biết điều gì đã xảy ra. Loại phân tích này giúp mô tả hoặc tóm tắt dữ liệu định lượng bằng cách trình bày các số liệu thống kê.
* Phân tích chuẩn đoán: cho biết vì sao những điều này lại xảy ra, loại phân tích này giúp hiểu sâu về nguyên nhân của những gì đã xảy ra trong quá khứ.
* Phân tích dự đoán: sử dụng dữ liệu để hình thành các dự báo về tương lai. Bằng những nguyên nhân đã được tìm ra ở phân tích chuẩn đoán có thể dự đoán điều gì sẽ xảy ra trong tương lai.
* Phân tích đánh giá: là bước phân tích khó nhất đòi hỏi kinh nghiệm và tầm nhìn khách quan, bởi vì sau 3 bước phân tích trên, người phân tích sẽ đưa ra những nhận xét và lời khuyên đề nghị cho doanh nghiệp,...

Ở lớp phân tích mô tả có thể thực hiện bằng Power BI. Tuy nhiên từ lới chuẩn đoán trở đi phải kết hợp hoặc sử dụng thuần thuật toán để đánh giá, nghiên cứu dữ liệu. Đối với lớp chuẩn đoán trở đi thì cần Machine Learning để phân tích.

CHƯƠNG 2 PHÂN TÍCH THIẾT KẾ

2.1 Đặt vấn đề

Những năm gần đây với sự phát triển không ngừng của trò chơi điện tử đã khiến cho lĩnh vực đầu tư này trở nên nóng bỏng hơn bao giờ hết. Trò chơi điện tử giờ đây không còn được nhắc đến với những lời lẻ tiêu cực, chỉ những thứ vô bổ, tốn thời gian. Trò chơi điện tử giờ đây đã trở thành một công cụ nơi mà nhà đầu tư dễ dàng thu lợi nhuận và người chơi có thể đem lại thu nhập cho mình từ việc chơi tựa game đó. Các nền tảng trực tuyến nổi tiếng hiện nay như Nimo TV, Facebook, Youtube cũng chú trọng nhiều đến mảng trò chơi, thông qua các nền tảng này người chơi có thể livestream để chia sẽ các khoảnh khắc của mình và kiếm tiền từ lược theo dõi và lược xem. Khi chỉ dừng lại ở đó trò chơi điện tử nói chung và thể thao điện tử nói riêng ngày càng được công nhận, điển hình như Seagame 31 lần đầu tiên trò chơi điện tử được đưa vào trở thành một bộ môn tranh tài như bao bộ môn khác. Góp mặt tại Seagame với 8 bộ môn và 10 nội dung thi đấu như: Liên minh huyển thoại tốc chiến, Liên minh huyền thoại, PUBG Mobile, Fifa Online 4, … đã cho thấy một bộ mặt khác trong lĩnh vực esport này.

Đầu tư trong lĩnh vữ trò chơi điện tử đã và đang cho thấy một bức tranh lợi nhuận to lớn, nhưng việc đầu tư và phát triển như thế nào và ở đâu thì lại là một bài toán nan giải. Để tạo ra một tựa game và phát triển nó để nhiều người biết đến và thu lại được lợi nhuận trên thực tế là điều không đơn giản. Có rất nhiều các dự án phát triển về game đã được triển khai những kết quả lại không được như mong đợi, khiến các nhà đầu tư thua lỗ lớn như: “Elemon”, ”Heroes Empires”, “Space Cryto”. Để có thể thành công trong việc xây dựng và phát hành một tựa game bên cạnh các công nghệ để xây dựng, lập trình nên nó cũng cần phải quan tâm đến việc nghiên cứu xu hướng phát triển của những tựa game qua các năm, tìm hiểu nền tảng hay thể loại nào đã và đang được yêu thích từ đó dễ dàng tiếp cận với người dùng và đem lại lợi nhuận cao.

2.2 Giải quyết vấn đề:

2.2.1 Scraping data

a) Giới thiệu công cụ:

- Python: ngôn ngữ lập trình bậc cao cho các mục đích lập trình đa năng được tạo ra năm 1991. Python được thiết kế với ưu điểm mạnh là dễ đọc, dễ học và dễ nhớ. Python là ngôn ngữ có hình thức sáng sủa, cấu trúc dễ rõ ràng, thuận tiện cho người mới học lập trình. Cấu trúc python cho phép người sử dụng viết mã lệnh với số lần gõ phím tối thiểu. Ban đầu Python được phát triển để chạy trên nên Unix. Dần theo thời gian mở rộng sang mọi hệ điều hành từ MS-DOS đến Mac OS, OS/2, Linux và các hệ điều hành khác thuộc họ Unix. Đặc biệt được sử dụng rỗng rãi trong lĩnh vực dữ liệu, phát triển trí tuệ nhân tạo (AI), trích xuất thông tin từ dữ liệu và máy học dạy máy tính tự động học hỏi từ dữ liệu và đưa ra các dự đoán chính xác. Có thể sử dụng Python cho các tác vụ khoa học dữ liệu sau:

+ Sửa và loại bỏ dữ liệu không chính xác, hay còn được gọi là làm sach dữ liệu

+ Trích xuất và chọn lọc các đặc điểm đa dạng của dữ liệu

+ Gắn nhãn dữ liệu, gán tên có ý nghĩa cho dữ liệu

+ Tìm các số liệu thống kê khác nhau từ dữ liệu

+ Trực quan hóa dữ liệu bằng cách sử dụng các biểu đồ và đồ thị, chẵn hạn như biểu đồ đường, biểu đồ cột, biểu đồ tần suất và biểu đồ tròn.

Sử dụng các thư việnc ML (Machine Learning) của Python để đào tạo các mô hình ML và xây dựng các công cụ phân loại giúp phân loại dữ liệu một cách chính xác. Những chuyên gia từ nhiều lĩnh vực sử dụng những công cụ phân loại dựa trên Python để thực hiện các tác vụ phân loại, chẵng hạn như phân loại hình ảnh, văn bản cũng như lưu lượng truy cập mạng, nhan dạng giọng nói và nhận diện khuân mặt. Các nhà khoa học dữ liệu cũng sử dụng Python cho deep learning, một kỹ thuật ML nâng cao.

* Visual Studio Code: là ứng dụng cho phép biên tập, soạn thảo các đoạn code để hỗ trợ trong quá trình thực hiện xây dựng và thiết kế chương trình một cách hiệu quả và nhanh chóng. Trình soạn thảo này hỗ trơ trên hầu hết các nền tảng như MacOS, Windows, Linux. Hơn thế nữa VS Code (Visual Studio Code) còn có khả năng tương thích với những máy tính có cấu hình tầm trung vẫn có thể sử dụng được dễ dàng.

Hỗ trợ đa dạng các chức năng Debug, đi kèm với Git, có Syntax Highlighting. Đặc biệt là tự hoàn thành mã thông minh, Snippets, và khả năng cải tiến mã nguồn. Nhờ tính năng tùy chỉnh, Visual Studio Code cũng cho phép người dùng thay đổi Theme, phím tắt,…

Visual Studio Code được chia thành 2 phiên bản Visual Studio Enterprise và Visual Studio Professional, các phiên bản cao cấp có tính phí được sử dụng bởi các công ty chuyên về lập trình. Bên cạnh đó Microsoft cũng cho ra mắt phiên bản Community của phần mềm, cung cấp cho người dùng những tíng năng cơ bản nhất, phù hợp với đối tượng lập trình không chuyên, mới tiếp cận với công nghệ

Sở hữu nhiều tính năng:

+ Đa nền tảng

+ Đa ngôn ngữ lập trình

+ Hỗ trợ website

+ Kho tiện ích mở rộng phong phú

+ Lưu trữ phân cấp

+ Màn hình đa nhiệm

+ Hỗ trợ Git

b) Thực hiện

Để giải quyết vấn đề phân tích xu hướng phát triển game như trên thực hiện scraping bằng ngôn ngữ Python trên trình soạn thảo Visual Studio Code.

Các bước thực hiện để scraping data:

* Khảo sát website, lựa chọn các trang web về game có thể request để lấy HTML
* Chọn các thông tin cần để scraping (tên, thể loại, thời gian, quốc gia, rate, vote, …)
* Phân tích cấu trúc HTML của trang web cần lấy thông tin
* Lựa chọn các thẻ chứa nổi dụng cần tìm
* Viết code để lấy thông tin qua ngôn ngữ Python

**2.2.Xử lý dữ liệu(ETL)**

Dữ liệu thô sau khi cào về vẫn còn bẩn(Null, Duplicate, Negative Values,…) dưới định dạng file .csv sẽ được đưa lên Azure Microsoft để tiến hành xử lý. Quy trình xử lý dữ liệu trên Azure sẽ qua các bước:

* Bước 1: Tạo Resource Group để nhóm các tài nguyên, tools.
* Bước 2: Tạo một Storage Account để chứa dữ liệu ở 3 level( tương ứng với 3 lớp dữ liệu là brown, silver, gold).
* Bước 4: Tạo một Data Brick để ETL dữ liệu.
* Bước 6: Tạo một SQL Database để chứa dữ liệu xử lý sạch(phục vụ cho mục đích visualize sau này).
* Bước 5: Tạo Data Factory để lấy dữ liệu ở container trong Storage Account vừa tạo để đưa vào Data Brick để xử lý và đẩy dữ liệu sạch về SQL DataBase.

**2.2.1. Resource Group**

Nhóm tài nguyên là một vùng chứa hợp lý liên kết nhiều tài nguyên giúp quản lý chúng như một thực thể duy nhất — dựa trên vòng đời và bảo mật. Khi khách hàng đưa ra yêu cầu quản lý một tài nguyên cụ thể. Trình quản lý tài nguyên Azure kết nối với nhà cung cấp tài nguyên cho loại tài nguyên đó để hoàn thành yêu cầu. Ví dụ: nếu khách hàng đưa ra yêu cầu quản lý tài nguyên máy ảo. Resource Group sẽ kết nối với Microsoft.Computenhà cung cấp tài nguyên. Resource Group gom các tài nguyên lại làm một, thuận tiện cho tìm kiếm, quản lý, bảo trì khi cần thiết.

**2.2.2. Storage Account**

Nền tảng Azure Storage là giải pháp lưu trữ đám mây của Microsoft dành cho các tình huống lưu trữ dữ liệu hiện đại. Dữ liệu trong Azure Storage có thể truy cập từ mọi nơi trên thế giới qua HTTP hoặc HTTPS. Microsoft cung cấp các thư viện máy khách cho Azure Storage bằng nhiều ngôn ngữ khác nhau, bao gồm .NET, Java, Node.js, Python, PHP, Ruby, Go và những thứ khác, cũng như API REST. Nền tảng Azure Storage bao gồm các dịch vụ dữ liệu sau:

* Azure Blob: Một kho lưu trữ đối tượng có thể mở rộng lớn cho dữ liệu văn bản và nhị phân. Cũng bao gồm hỗ trợ phân tích dữ liệu lớn thông qua Data Lake Storage Gen2.
* Azure File: Chia sẻ tệp được quản lý để triển khai trên đám mây hoặc tại chỗ.
* Azure Queues: Một cửa hàng nhắn tin để gửi tin nhắn đáng tin cậy giữa các thành phần ứng dụng.
* Azure Tables: Một cửa hàng NoSQL để lưu trữ dữ liệu có cấu trúc mà không cần toán học.
* Azure Disks: Khối lượng lưu trữ cấp khối cho máy ảo Azure.

Trong dự án lần này, nhóm sẽ tập trung chủ yếu khai thác tính năng và cách sử dụng Azure Blob để tạo các container chưa dữ liệu ở 3 level(Level 1 chứa dữ liệu thô, level 2 chứa dữ liệu xử lý và level 3 chứa dữ liệu sạch). Bộ nhớ Blob được tối ưu hóa để lưu trữ một lượng lớn dữ liệu phi cấu trúc, chẳng hạn như văn bản hoặc dữ liệu nhị phân. Bộ nhớ Blob là nơi lưu trữ lý tưởng của:

* Hình ảnh hoặc tài liệu trực tiếp đến trình duyệt.
* Tệp dữ liệu để truy cập phân tán.
* Truyền phát video và âm thanh.
* Lưu trữ dữ liệu để sao lưu và khôi phục, khôi phục sau thảm họa và lưu trữ.
* Lưu trữ dữ liệu để phân tích bằng dịch vụ tại chỗ hoặc dịch vụ lưu trữ trên Azure.

Các đối tượng trong bộ lưu trữ Blob có thể được truy cập từ mọi nơi trên thế giới thông qua HTTP hoặc HTTPS. Người dùng hoặc ứng dụng khách có thể truy cập các đốm màu thông qua URL, API Azure Storage REST , Azure PowerShell , Azure CLI hoặc thư viện ứng dụng Azure Storage.

**2.2.3. Data Brick**

Databricks là việc triển khai Apache Spark trên Azure. Với các cụm Spark được quản lý hoàn toàn, nó được sử dụng để xử lý khối lượng công việc lớn của dữ liệu và cũng hỗ trợ trong kỹ thuật dữ liệu, khám phá dữ liệu và cũng như trực quan hóa dữ liệu bằng Máy học.  Ngoài hỗ trợ nhiều ngôn ngữ, dịch vụ này cho phép tích hợp dễ dàng với nhiều dịch vụ Azure như Blob Storage, Data Lake Store, SQL Database và các công cụ BI như Power BI, Tableau,…

Databricks trên nền tảng Azure thân thiện hỗ trợ người dùng nhiều tính năng mà như sau:

* Truy cập Azure Blob Storage từ Azure Databricks
* Xử lý và khám phá dữ liệu trong Azure Databricks
* Kết nối Cơ sở dữ liệu Azure SQL với Cơ sở dữ liệu Azure
* Tải dữ liệu vào Kho dữ liệu Azure SQL.
* Tích hợp với Power BI để visualize dữ liệu.
* Chạy Azure Databricks Notebook trong Azure Data Factory,…

Như đã nêu ở chương 1, ngôn ngữ mà nhóm sử dụng để tiến hành xử lý dữ liệu chính là Scala. Mặc dù là ngôn ngữ mới, nhưng scala có nhiều tính năng vượt trội thích hợp để tiến hành xử lý dữ liệu – đặc biệt là big data như: hiệu xuất cao(gấp 10 lần so với Python, nhanh hơn 20% so với Java), dễ nhớ dễ sử dụng và đặc biệt được Databricks hỗ trợ nhiều thư viện.

**2.2.4. SQL Database**

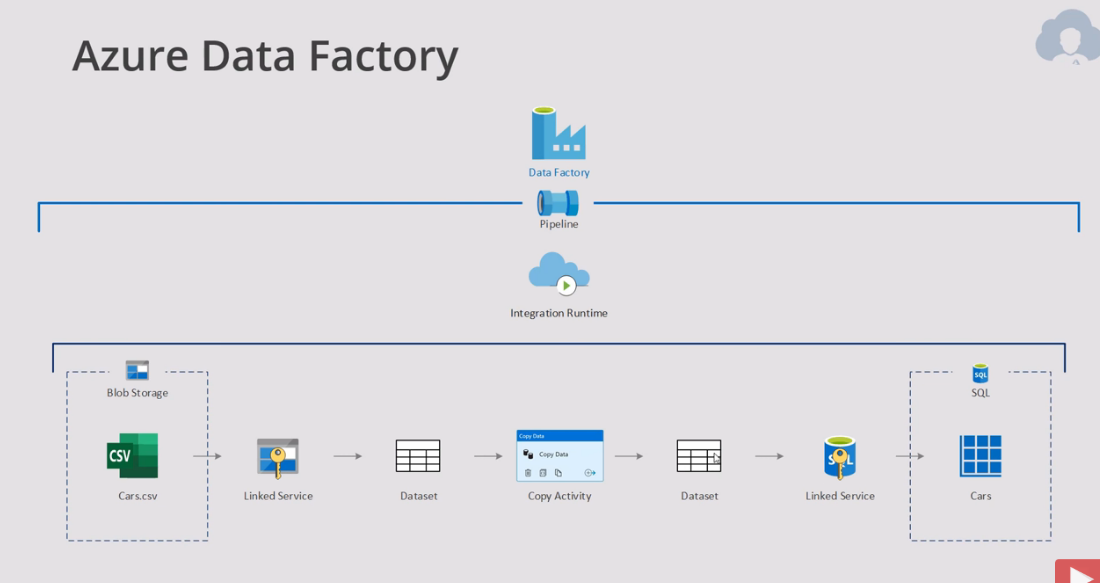
Về cơ bản SQL Database triển khai trên nền tảng Azure có các chức năng giống với SQL Server được cài đặt trên máy local. Tuy nhiên, SQL Database trên Cloud có tốc độ xử lý nhanh hơn, chứa được lượng lớn dữ liệu, không bị phụ thuộc vào cấu hình mấy vật lý. Ngoài ra, Azure cũng thiết kế trình quản lý SQL Database thân thiện dễ dàng sử dụng.

SQL Database sau khi tạo sẽ là nơi chứa dữ liệu sạch, trong SQL Database còn tích hợp Power BI hỗ trợ người dùng dễ dàng visualize dữ liệu hơn. Đảm bảo mọi thao tác đều có thể thực hiện trên Cloud. Dữ liệu sau khi xử lý xong tại Databricks sẽ được Data Factory chuyển đổi vào SQL Database thông qua các Pipeline.

Trong dự án lần này nhóm sẽ tiến hành chứa dữ liệu đã sạch và có ý nghĩa tại 2 nơi: một nơi là Container của Storange Account để lưu trữ và một nơi là SQL Database để tiến hành truy vấn dữ liệu, visualize cũng như chỉnh sửa nếu cần thiết.

**2.2.5. Data Factory**

Phương pháp quản lý dữ liệu đã phát triển qua nhiều năm với nhiều lần lặp lại như Xử lý giao dịch trực tuyến (OLTP), Kho dữ liệu và Marts dữ liệu, Xử lý phân tích trực tuyến (OLAP), Hồ dữ liệu và cuối cùng là khái niệm Data Lakehouse đang trở nên phổ biến trong kỷ nguyên đám mây. Khi quy mô dữ liệu phát triển và lược đồ của các đối tượng dữ liệu thay đổi theo thời gian, việc phù hợp với tốc độ phát triển các đường ống dữ liệu điểm-điểm mới cũng như duy trì các đường ống dữ liệu hiện có ngày càng trở nên khó khăn và kém hiệu quả.



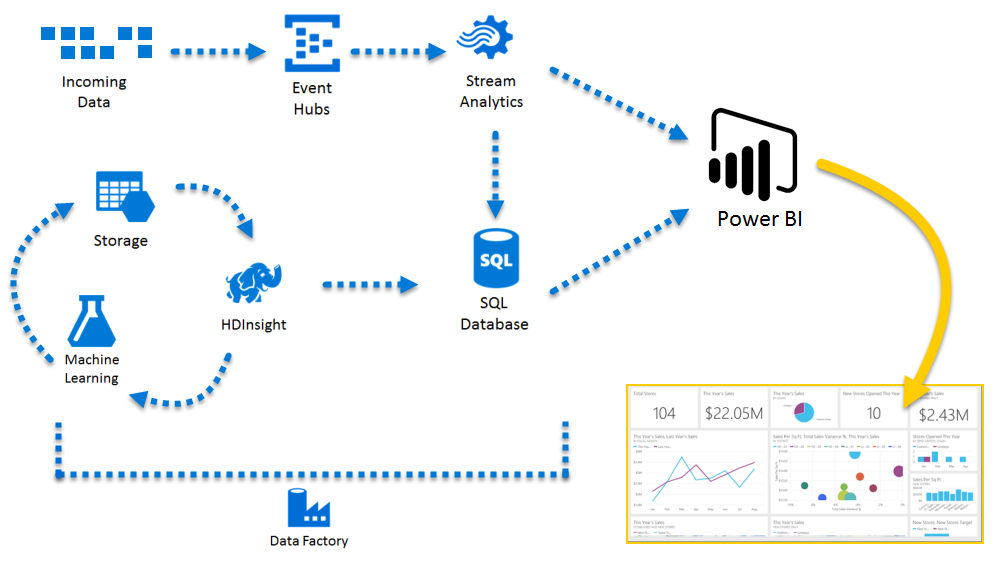
*Hình. Tiến trình xử lý dữ liệu tại ADF*

Data Factory chứa một loạt các hệ thống được kết nối với nhau cung cấp một nền tảng end-to-end hoàn chỉnh cho các kỹ sư dữ liệu. Azure Data Factory bao gồm các thành phần chính dưới đây:

* Pipelines: một nhóm hợp lý các hoạt động thực hiện một đơn vị công việc.
* Activities: Tạo và quản lý đồ thị logic chuyển đổi dữ liệu mà người dùng có thể sử dụng để chuyển đổi dữ liệu có kích thước bất kỳ.
* Datasets: Tập dữ liệu đại diện cho cấu trúc dữ liệu trong kho dữ liệu, trỏ đến hoặc tham chiếu dữ liệu người dùng muốn sử dụng trong các hoạt động của mình dưới dạng đầu vào hoặc đầu ra.
* Linked services: tạo liên kết xác định kết nối với nguồn dữ liệu và một tập dữ liệu đại diện cho cấu trúc của dữ liệu.
* Data Flows
* Integration Runtimes

**2.3. Trực quan hoá dữ liệu bằng Power BI**

Dữ liệu sau khi làm sạch và được đổ vào SQL Database sẽ được tiến hành visualize thành nhiều biểu đồ khác nhau. Power BI tích hợp trong SQL Server hỗ trợ người dùng dễ dàng lấy dữ liệu thông qua các câu lệnh Select. Nhờ đố việc trực quan diễn ra chính xác hơn.



*Hình. Tiến trình tạo ra bảng dashboard trong Power BI*

Sử dụng Power BI trên Azure mang lại nhiều lợi ích vượt trội:

* Bắt đầu nhanh chóng mà không cần thiết lập cơ sở hạ tầng
* Kết hợp và phân tích dữ liệu từ nhiều nguồn
* Dễ dàng xác định các chỉ số và KPI chính
* Dễ dàng mở rộng quy mô tài nguyên để đáp ứng nhu cầu thay đổi
* Bảo mật dữ liệu trong một mô hình dữ liệu duy nhất, đáng tin cậy
* Xử lý khối lượng lớn dữ liệu trong thời gian nhanh nhất
* Tích hợp và quản lý giải pháp thông minh kinh doanh của bạn một cách an toàn

# 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 4.1 Kiến trúc**

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

**Hình 4.1 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