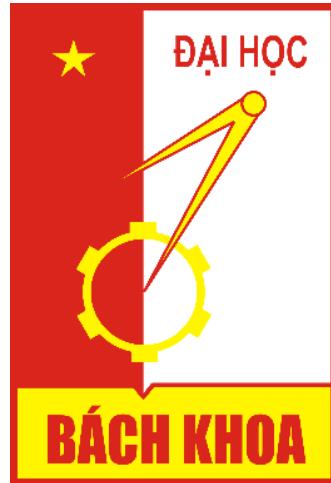


TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI  
VIỆN TOÁN ỨNG DỤNG VÀ TIN HỌC

----- ☐ & ☐ -----



**BÁO CÁO GIỮA KÌ  
MÔN KHO DỮ LIỆU VÀ KINH DOANH THÔNG MINH**

**Nhóm 18**  
**Đề tài : Education**

**Giáo viên hướng dẫn:** Thầy Nguyễn Danh Tú

**Nhóm sinh viên thực hiện:**

Nguyễn Tuấn Hùng	- 20206284
Trần Hữu Tiên	- 20206264
Bùi Hồng Giang	- 20206280
Trịnh Vũ Thiên	- 20206306

**Đánh giá nhóm**

Phân chia công việc	Hùng	Thiên	Giang	Tiên
Phân I	x		x	
Phân II	x	x		x
Phân III		x		x

Phân IV			x	
Làm Slide	x	x	x	
Thuyết trình	x	x		x
Làm báo cáo	x		x	x

BẢNG ĐÁNH GIÁ THÀNH VIÊN						
MÔN HỌC: Kho dữ liệu và kinh doanh thông minh						
	HỌ VÀ TÊN:	Trần Hữu Tiên				
	LỚP:	K65 – Hệ thống thông tin quản lý 02				
	NHÓM	N18				
STT	Tên thành viên	Làm tốt phần việc được giao	Liên hệ được khi cần	Khả năng đóng góp sáng kiến, ý kiến cho hoạt động nhóm	Sẵn sàng giúp đỡ	Đóng góp chung vào kết quả của nhóm
1	Trần Hữu Tiên	5	4	5	4	5
2	Nguyễn Tuấn Hùng	4	5	4	5	5
3	Bùi Hồng Giang	4	5	5	5	4
4	Trịnh Vũ Thiên	5	5	4	5	4

## Mục lục

Đánh giá nhóm .....	0
<b>Mục lục .....</b>	<b>2</b>
LỜI MỞ ĐẦU .....	<b>Error! Bookmark not defined.</b>
LỜI MỞ ĐẦU .....	5
CHƯƠNG 1: TỔNG QUAN VỀ DATA WAREHOUSE .....	6
1. Data Warehouse là gì ? .....	6
2. Các khái niệm cơ bản .....	6
3. Đặt tính của Data Warehouse .....	8
4. Kiến trúc của Data Warehouse .....	9
5. Lợi ích của Data Warehouse .....	9
CHƯƠNG 2: TỔNG QUAN VỀ BUSINESS INTELLIGENCE .....	11
1. Business Intelligence là gì ? .....	11
2. Các thành phần cơ bản của hệ thống Business Intelligence .....	11
3. Quy trình của Business Intelligence .....	12
4. Lợi ích của Business Intelligence .....	13
CHƯƠNG 3: ỨNG DỤNG DATA WAREHOUSE VÀ BI VÀO GIÁO DỤC .....	13
<b>I. Khảo sát .....</b>	<b>13</b>
1. Tổng quan .....	13
- Quản lý thông tin trong lĩnh vực giáo dục là quá trình thu thập, lưu trữ, tổ chức và xử lý thông tin liên quan đến hệ thống giáo dục .....	13
- Nó bao gồm việc quản lý thông tin về học sinh, sinh viên, giáo viên, chương trình học, học phí và đánh giá trường học, giáo viên, học sinh trong lĩnh vực giáo dục, ... .....	13
- Tại sao nên sử dụng ? .....	13
• Tăng cường khả năng ra quyết định .....	13
• Nâng cao hiệu suất và chất lượng giáo dục .....	14
• Tối ưu hóa tài nguyên .....	14
- Chức năng .....	14
• Thu thập thông tin .....	14
• Lưu trữ và tổ chức thông tin .....	14
• Xử lý và phân tích thông tin .....	14
• Bảo mật thông tin .....	14

• Chia sẻ thông tin .....	14
- Thực trạng : trên thực tế, nhiều trường học và cơ sở giáo dục ở Việt Nam đã bắt đầu sử dụng hệ thống quản lý thông tin trong giáo dục. Tuy nhiên, việc triển khai và sử dụng quản lý thông tin giáo dục vẫn chưa đồng nhất và phổ biến.....	14
2. Quy trình nghiệp vụ .....	14
- Quy trình xử lý học tập : .....	14
- Quy trình xử lý học phí : .....	15
- ER Diagram : .....	15
3. Quy mô dữ liệu.....	16
- Về chi tiết các bảng dữ liệu :.....	16
- Các file dữ liệu là thông tin dữ liệu giáo dục tại các trường học tại thành phố Hà Nội .....	16
- Kích thước : 163 MB .....	16
4. Yêu cầu phân tích.....	16
Chúng em chọn phân tích bộ dữ liệu theo 3 mặt sau .....	16
- Scores : .....	16
• Phân tích điểm số trung bình.....	16
• Điểm số theo môn học.....	16
• Phân loại chất lượng học sinh .....	16
- Attendance : .....	16
• Phân tích số buổi vắng mặt của học sinh .....	16
• Phân tích theo lớp học , trường học .....	16
• Phân tích tương quan giữa số buổi nghỉ học và điểm số .....	16
- Tuition fee : .....	16
• Phân tích tổng quan.....	16
• Học phí theo từng lớp học .....	16
• Học phí theo trường học.....	16
• Phân tích tương quan giữa học phí và điểm số .....	16
<b>II. Phân tích thiết kế.....</b>	16
1. Data Exploration .....	16
- Bảng Attendance : .....	16
.....	17
Biểu đồ histogram thể hiện phân phối số lượng buổi nghỉ học .....	17
.....	17
4. Xây dựng data model OLAP .....	31

-	Data model logic .....	31
-	Data model vật lý .....	32
5.	Xây dựng cơ sở dữ liệu OLAP .....	32
-	Xây dựng qua OLAP Views .....	32
•	Tạo view Address.....	32
		32

## LỜI MỞ ĐẦU

Trong thời đại kỹ thuật số hiện nay, dữ liệu là tài nguyên quan trọng và có giá trị lớn đối với các doanh nghiệp, tổ chức và cá nhân. Với số lượng dữ liệu khổng lồ được tạo ra hàng ngày, việc phân tích và trích xuất thông tin từ dữ liệu đã trở thành một thách thức lớn, điều đó đòi hỏi sự ra đời của các hệ thống hỗ trợ nghiệp vụ phân tích, trích xuất và tạo báo cáo. Từ đó cho nhiều góc nhìn hơn về dữ liệu giúp cho những đánh giá, quyết định của người sử dụng trở nên đúng đắn hơn.

Trong đồ án này, em sẽ trình bày về cách thiết kế và xây dựng kho dữ liệu để giúp tổ chức đáp ứng nhu cầu phân tích dữ liệu một cách nhanh chóng và hiệu quả. Kho này sẽ được xây dựng trên nền tảng các công nghệ phần mềm và phần cứng tiên tiến nhất để đảm bảo tính tin cậy và hiệu suất cao.

Nội dung của đồ án này bao gồm 3 chương chính với mục tiêu giúp chỉ ra các hướng phân tích của một dữ liệu, nêu ra các cơ sở lý thuyết cần có để có thể xây dựng một kho dữ liệu. Sau là các bước làm sạch dữ liệu, xác định chiều sâu của dữ liệu từ đó chuyển đổi dữ liệu qua các cơ sở dữ liệu OLTP, OLAP. Từ đó sử dụng các công cụ phân tích xây dựng các báo cáo, phân tích về dữ liệu.

### Tóm tắt nội dung

#### Chương 1: Tổng quan về Data Warehouse

- Giới thiệu chung về dữ liệu, các khái niệm về data warehouse, OLTP, OLAP.
- Kiến trúc của data warehouse.
- Các đặc tính, lợi ích của data warehouse.

#### Chương 2: Tổng quan về Business Intelligence

- Giới thiệu khái niệm, các thành phần cơ bản của hệ thống BI.
- Quy trình, lợi ích của BI.

#### Chương 3: Ứng dụng Data Warehouse và Business Intelligence vào bài toán giáo dục

- Phân tích các kiến trúc, luồng dữ liệu cũ từ đó đưa ra gợi ý cho hệ thống mới.
- Phân tích, đưa ra các kiến trúc của hệ thống, các bước xử lý dữ liệu nguồn, phân tích dữ liệu theo các chiều và đưa ra phương hướng thiết kế.
- Truyền dữ liệu vào mô hình, xây dựng báo cáo trên công cụ Power BI.

Trước khi đi vào bài báo cáo, lời đầu tiên cho em xin chân thành cảm ơn sự hỗ trợ và hướng dẫn của **ThS. Nguyễn Danh Tú**, giảng viên Bộ môn Toán Tin tại viện Toán Ứng Dụng và Tin Học, Đại học Bách Khoa Hà Nội. Cảm ơn thầy đã luôn tận tâm tư vấn, hướng dẫn và luôn sẵn sàng hỗ trợ em trong quá trình thực hiện đồ án này. Qua quá trình làm bài tập lóng chung em đã nhận ra được nhiều điều mới mẻ và rất nhiều kỹ năng có ích cho hướng đi của em sau này. Tất cả những điều đó đã góp phần giúp đỡ em hoàn thành được đồ án theo ý tưởng ban đầu và đáp ứng được những yêu cầu đặt ra.

Chúng em xin chân thành cảm ơn.

## CHƯƠNG 1: TỔNG QUAN VỀ DATA WAREHOUSE

### 1. Data Warehouse là gì ?

Data Warehouse có nghĩa là kho dữ liệu là một loại quản lý dữ liệu hệ thống được thiết kế để cho phép và hỗ trợ kinh doanh thông minh (BI), đặc biệt là phân tích.

- + Data Warehouse chỉ nhằm mục đích thực hiện các truy vấn, phân tích và thường chứa một lượng lớn dữ liệu.
- + Khả năng phân tích Data Warehouse cho phép các tổ chức thu được những hiểu biết kinh doanh có giá trị từ dữ liệu của họ để cải thiện việc ra quyết định. Theo thời gian, nó xây dựng một hồ sơ lịch sử có thể là vô giá đối với các nhà Data Science và nhà phân tích kinh doanh.

### 2. Các khái niệm cơ bản

#### 2.1. OLTP

Hệ thống OLTP (Online Transaction Processing – Xử lý giao dịch trực tuyến):

- + Dữ liệu phát sinh từ các hoạt động hằng ngày.
- + Thu thập xử lý để phục vụ công việc nghiệp vụ cụ thể của một tổ chức.
- + Thường được gọi là dữ liệu tác vụ và hoạt động thu thập xử lý dữ liệu này.

#### 2.2. Data Warehouse

- Kho dữ liệu phục vụ cho việc phân tích với kết quả mang tính thông tin cao.
- Kho dữ liệu là nơi dữ liệu được tuyển tập và lưu trữ:

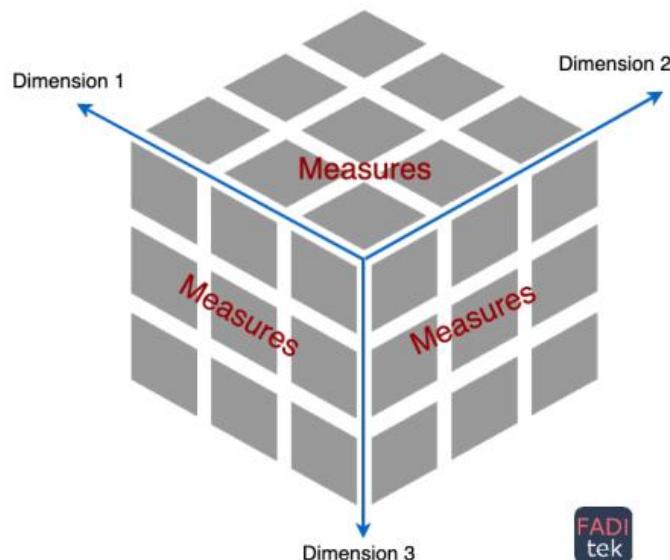
- + Hướng chủ đề
- + Tích hợp
- + Biến đổi theo thời gian
- + Ôn định

- Kho dữ liệu dùng để hỗ trợ ra quyết định trong quản lý.

#### 2.3. OLAP

Hệ thống OLAP(On-Line Analytical Processing – Xử lý phân tích trực tuyến):

Hình ảnh sau đây là một data cube, trong bối cảnh phân tích dữ liệu và kho dữ liệu, là một biểu diễn dữ liệu đa chiều cho phép phân tích hiệu quả và linh hoạt.



- + OLAP là kỹ thuật sử dụng các cube (khối – thể hiện dữ liệu đa chiều) nhằm cung cấp khả năng truy xuất nhanh đến dữ liệu của kho dữ liệu. Tạo khối (cube) cho dữ

liệu trong các bảng chiều (dimension table) và bảng sự kiện (fact table) và cung cấp khả năng thực hiện các truy vấn tinh vi và phân tích cho các ứng dụng client.

- + OLAP là kỹ thuật cho phép các ứng dụng client truy xuất hiệu quả dữ liệu này.
- + OLAP được đặt ra để xử lý các truy vấn liên quan đến lượng dữ liệu rất lớn mà nếu cho thực thi các truy vấn này trong hệ thống OLTP sẽ không thể cho kết quả hoặc sẽ mất rất nhiều thời gian. OLAP - Online analytical processing- xử lý phân tích trực tuyến là một phần mềm tính toán cho phép người dùng trích xuất và truy vấn dữ liệu một cách chọn lọc và dễ dàng. OLAP là dữ liệu đa chiều, điều này có nghĩa là tất cả thông tin có thể được so sánh theo nhiều cách thức khác nhau.

Hệ thống OLAP được phân thành 3 loại chính:

- + MOLAP: OLAP hoạt động trực tiếp với khối OLAP đa chiều được gọi là OLAP đa chiều hay MOLAP. Đối với hầu hết mọi mục đích sử dụng, MOLAP là loại phân tích dữ liệu đa chiều nhanh nhất và thiết thực nhất.
- + ROLAP: ROLAP hay OLAP quan hệ, là loại phân tích trực tuyến (OLAP) các mô hình dữ liệu đa chiều. Điểm khác biệt giữa các ROLAP và MOLAP là nó truy cập dữ liệu lưu trữ ngay trong cơ sở dữ liệu quan hệ thay vì dữ liệu cơ sở đa chiều (cơ sở dữ liệu được sử dụng phổ biến trong các OLAP).Thêm vào đó, nó cũng có thể tạo ra các truy vấn SQL với mục đích thực hiện các phép tính khi người dùng cuối muốn như vậy.
- + HOLAP: HOLAP hay OLAP kết hợp là sự kết hợp giữa ROLAP (xử lý phân tích trực tuyến quan hệ) và MOLAP (xử lý phân tích trực tuyến đa chiều). HOLAP cung cấp lợi thế từ cả hai quy trình ROLAP và MOLAP do nó hỗ trợ cho cả hai định dạng lưu trữ. HOLAP để giải quyết câu hỏi :" cái nào tốt hơn?" bằng việc kết hợp khả năng xử lý của MOLAP và dung lượng dữ liệu của MOLAP. Ngoài ra, do có kiến trúc phức tạp và phải lưu trữ, xử lý tất cả cơ sở dữ liệu từ MOLAP và ROLAP nên HOLAP sẽ yêu cầu cập nhật và bảo trì thường xuyên hơn.

Mô hình chiều dữ liệu:

- + Dimension: Các bảng dimension được sử dụng để mô tả dữ liệu mà chúng ta muốn lưu trữ.  
Ví dụ: một nhà bán lẻ muốn lưu trữ thời gian, cửa hàng, và nhân viên tham gia vào một hóa đơn. Mỗi một bảng dimension là một danh mục của chính nó (ngày tháng, nhân viên, cửa hàng) và có thể có một hoặc nhiều thuộc tính (attributes). Với mỗi một cửa hàng, chúng ta lưu chúng các thông tin như vị trí trong thành phố, vùng miền, tỉnh thành và quốc gia. Mỗi một ngày tháng chúng ta lưu năm, tháng, ngày trong tháng, ngày trong tuần...Điều này liên quan đến sự phân cấp của các thuộc tính trong bảng dimension.
- + Fact: Bảng Fact chứa dữ liệu mà chúng ta muốn thêm vào reports, tổng hợp trên các giá trị trong các bảng dimension. Một bảng fact chỉ có các cột lưu giá trị và các cột khoá ngoại tham chiếu đến bảng dimensions. Kết hợp tất cả các khoá ngoại và khoá chính trong bảng fact.

Một số mô hình đơn giản:

- + Lược đồ bông tuyết (Snowflake schema):
  - Ưu điểm
    - Kích thước các bảng Dimension giảm, nâng cao tốc độ truy vấn.
    - Cho phép thực hiện các truy vấn phức tạp theo các chiều.

- Dễ dàng thiết lập và bảo trì.
- Nhược điểm:
  - Số lượng các bảng cần quản lý tăng.
  - Truy vấn xâu cần sự kết nối giữa nhiều bảng gây giảm hiệu năng.
- + Lược đồ hình sao (Star schema):
  - Ưu điểm:
    - Cải tiến hiệu năng, thời gian truy vấn nhanh.
    - Có ít bảng và cấu trúc đơn giản.
  - Nhược điểm:
    - Do có ít bảng nên lượng thông tin trong bảng là lớn, gây dư thừa.
    - Các mối quan hệ không được thể hiện quá rõ ràng.

#### 2.4. So sánh OLAP và OLTP

Yếu tố	OLAP	OLTP
Người dùng	<ul style="list-style-type: none"> <li>• Người ra quyết định</li> <li>• Nhóm người quản lý được chỉ định</li> </ul>	<ul style="list-style-type: none"> <li>• Nhóm nhân viên, khách hàng</li> <li>• Có thể truy cập bởi rất nhiều người</li> </ul>
Chức năng	<ul style="list-style-type: none"> <li>• Phân tích đa chiều, khai thác dữ liệu</li> <li>• Phân tích phức tạp, báo cáo kinh doanh</li> </ul>	<ul style="list-style-type: none"> <li>• Xử lý giao dịch gần đây</li> <li>• Nhanh chóng và chính xác</li> </ul>
Bản chất	<ul style="list-style-type: none"> <li>• Cung cấp thông tin tóm tắt, tổng hợp cho người truy vấn dữ liệu</li> </ul>	<ul style="list-style-type: none"> <li>• Ghi lại các bước thao tác</li> </ul>
Thiết kế	<ul style="list-style-type: none"> <li>• Theo hướng chủ đề và xem thông tin dưới dạng đa chiều</li> </ul>	<ul style="list-style-type: none"> <li>• Theo hướng ứng dụng và xem bản ghi dưới dạng tập hợp bảng</li> </ul>
Dữ liệu	<ul style="list-style-type: none"> <li>• Cần thông tin dữ liệu trong vài năm</li> </ul>	<ul style="list-style-type: none"> <li>• Cần trạng thái hiện tại của dữ liệu</li> </ul>
Loại sử dụng	<ul style="list-style-type: none"> <li>• Thường ít được cập nhật dữ liệu</li> </ul>	<ul style="list-style-type: none"> <li>• Thao tác đọc và ghi</li> </ul>
Chế độ xem	<ul style="list-style-type: none"> <li>• Tổng hợp dữ liệu từ nhiều nguồn, trong quá khứ</li> </ul>	<ul style="list-style-type: none"> <li>• Tập trung vào dữ liệu hiện tại của tổ chức</li> </ul>
Các mẫu truy cập	<ul style="list-style-type: none"> <li>• Chủ yếu là các hoạt động đọc, nghiên cứu</li> </ul>	<ul style="list-style-type: none"> <li>• Các giao dịch ngắn</li> </ul>

### 3. Đặt tính của Data Warehouse

- Hướng chủ đề
  - + Kho dữ liệu được thiết kế để hỗ trợ trong việc phân tích dữ liệu.
  - + Được tổ chức xung quanh các chủ đề chính như: khách hàng, sản phẩm, bán hàng,...
  - + Loại bỏ những dữ liệu không hữu ích cho trình ra quyết định.
- Tích hợp
  - + Là đặc tính quan trọng nhất của kho dữ liệu.
  - + Dữ liệu được tập hợp từ nhiều nguồn khác nhau: Cơ sở dữ liệu quan hệ (relational databases), flat files, các bảng ghi toàn tác trực tuyến. Điều này sẽ dẫn đến việc

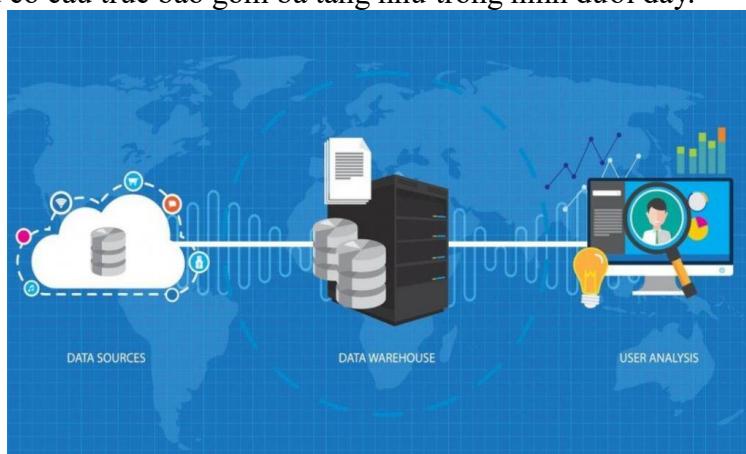
trong quá trình tập hợp dữ liệu phải thực hiện việc làm sạch, sắp xếp, rút gọn dữ liệu.

- Ôn định
  - + Được lấy từ nhiều nguồn dữ liệu của hệ thống tác nghiệp có sẵn.
  - + Kho dữ liệu tách rời vật lý với môi trường tác nghiệp, nên dữ liệu trong kho dữ liệu là dữ liệu chỉ đọc, không chỉnh sửa hoặc thêm mới được.
- Biến đổi theo thời gian
  - + Dữ liệu quá khứ và hiện tại.
  - + Mỗi dữ liệu trong kho dữ liệu đều được gắn với thời gian và có tính lịch sử.

**Chú ý: Dữ liệu trong kho dữ liệu rất lớn và không được thêm, xóa, sửa dữ liệu.**

#### 4. Kiến trúc của Data Warehouse

Kho dữ liệu có cấu trúc bao gồm ba tầng như trong hình dưới đây.



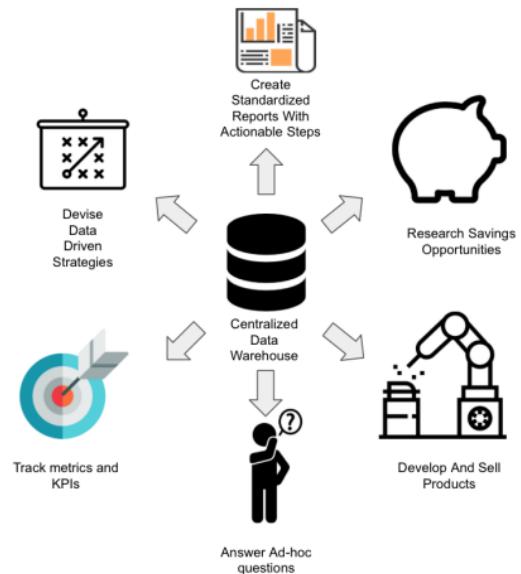
- + Tầng dữ liệu đầu vào (Data Sources): là nơi chứa các dữ liệu của doanh nghiệp. Dữ liệu này có thể là: website bán hàng, phần mềm kế toán, quản lý nhân sự, quản lý khách hàng (CRM), hệ thống lõi ngân hàng (Corebanking), hệ thống thẻ, hệ thống quản lý thanh toán online, . . .
- + Tầng giữa (Data Warehouse): là nơi thu thập, tích hợp dữ liệu từ nhiều nguồn khác nhau sau đó chuẩn hóa về cùng định dạng, làm sạch xử lý dữ liệu để tìm lỗi và sửa và lưu trữ dữ liệu đã tổng hợp.
- + Tầng phân tích dữ liệu (User Analysis): nơi thực hiện các thao tác truy vấn, báo cáo, phân tích để tìm ra xu hướng, trung bình, tổng hợp, . . .

#### 5. Lợi ích của Data Warehouse

- Các vấn đề:
  - + Đối với các tổ chức có lượng dữ liệu ngày càng lớn thì càng khó truy cập và sử dụng dữ liệu.
  - + Dữ liệu trong nhiều định dạng khác nhau, tồn tại trên nhiều nền tảng khác nhau, và lưu trữ trong nhiều tập tin khác nhau, cấu trúc cơ sở dữ liệu khác nhau được phát triển bởi các nhà cung cấp khác nhau.
  - + Tổ chức phải viết và duy trì hàng trăm chương trình để trích xuất, chuẩn bị, hợp nhất dữ liệu để sử dụng cho nhiều chương trình khác nhau dùng để phân tích và báo cáo.
  - + Người ra quyết định muốn khai thác sâu hơn vào các dữ liệu.

- + Điều này dẫn đến các yêu cầu phát triển chương trình trích xuất mới hơn. Quá trình này rất tốn kém, không hiệu quả và tốn thời gian.

Sau đây là hình ảnh về 6 lợi ích của Data Warehouse



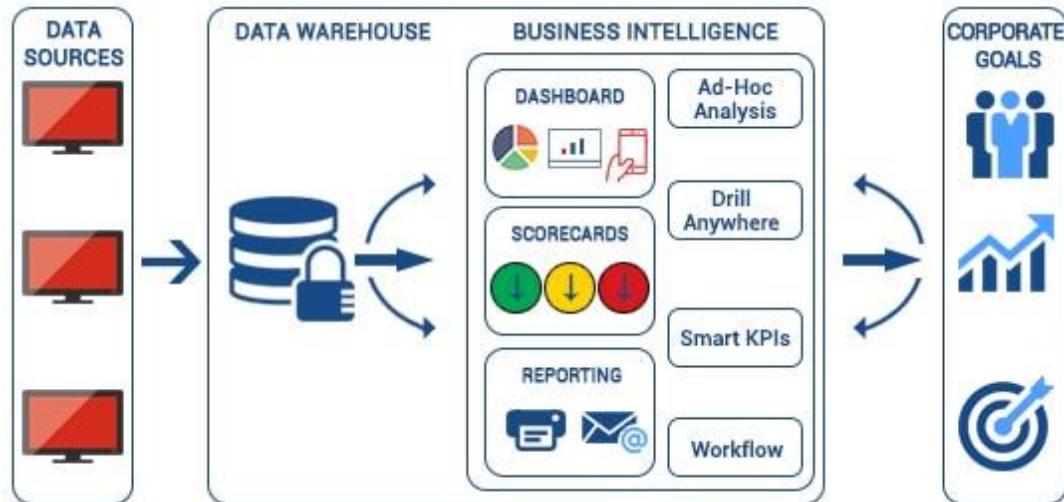
- Data Warehouse sẽ cung cấp một phương pháp tiếp cận tốt hơn.
  - + Data Warehouse thực hiện quá trình truy cập dữ liệu từ các nguồn không đồng nhất; làm sạch, lọc và chuyển đổi dữ liệu; lưu trữ dữ liệu theo cấu trúc để dễ dàng truy cập, hiểu rõ và sử dụng.
  - + Duy trì lịch sử dữ liệu, ngay cả khi các hệ thống giao dịch nguồn không.
  - + Trình bày thông tin của tổ chức một cách nhất quán.
  - + Cơ cấu lại dữ liệu để nó mang lại hiệu suất truy vấn tuyệt vời, ngay cả đối với các truy vấn phân tích phức tạp, mà không ảnh hưởng đến các hệ thống hoạt động.
  - + Đưa ra quyết định hỗ trợ truy vấn dễ dàng hơn để viết.

## CHƯƠNG 2: TỔNG QUAN VỀ BUSINESS INTELLIGENCE

### 1. Business Intelligence là gì ?

Business Intelligence là chuỗi những quy trình, kiến trúc, công nghệ giúp chuyển đổi dữ liệu thô thành thông tin có ý nghĩa. Từ đó, chúng giúp doanh nghiệp nhanh chóng đưa ra chiến lược hiệu quả nhất nhằm thu lại lợi nhuận cao.

Hình ảnh dưới là mô hình Business Intelligence.



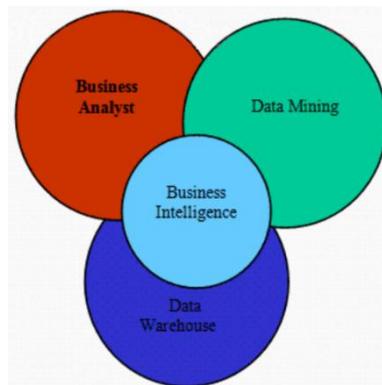
Những thông tin được tổng hợp từ BI mang đến một cái nhìn trực quan về doanh nghiệp. Đồng thời BI giúp doanh nghiệp xác định được những xu hướng và vấn đề hiện tại của thị trường. Về cơ bản, hệ thống Business Intelligence là hệ thống hỗ trợ quyết định dựa trên dữ liệu (Decision Support Systems – DSS). Business Intelligence cung cấp quan điểm lịch sử, hiện tại và dự đoán về hoạt động kinh doanh. Các chức năng phổ biến của BI bao gồm:

- + Xử lý phân tích trực tuyến (Online analytical processing)
- + Khai thác dữ liệu (Data mining)
- + Hỗ trợ quyết định (Decision support)
- + Truy vấn và báo cáo (Query and reporting)
- + Phân tích thống kê (Statistical analysis)
- + Phân tích dự đoán và phân tích theo quy định (Forecasting)

### 2. Các thành phần cơ bản của hệ thống Business Intelligence

Vấn đề cốt lõi trong hệ thống BI là kho dữ liệu (Data Warehouse) và khai phá dữ liệu (Data Mining) vì dữ liệu dùng trong BI là dữ liệu tổng hợp (nhiều nguồn, nhiều định dạng, phân tán và có tính lịch sử) đó chính là đặc trưng của kho dữ liệu. Đồng thời việc phân tích dữ liệu trong Business Intelligence không phải là những phân tích đơn giản (Query, Filtering) mà là những kỹ thuật trong khai phá dữ liệu (Data Mining) dùng để phân loại (classification) phân cụm (Clustering), hay dự đoán (Prediction). Vì vậy BI có mối quan hệ rất chặt chẽ với Data Warehouse và Data Mining.

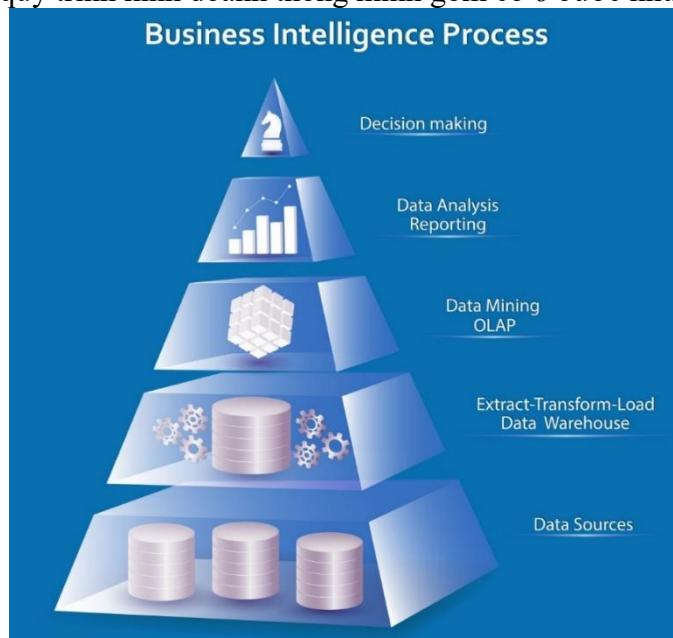
Hệ thống Business intelligence đơn giản có thể được xem là sự kết hợp của 3 thành phần chính như ảnh minh họa sau:



- + Data Warehouse (Kho dữ liệu): Chứa dữ liệu tổng hợp của doanh nghiệp.
- + Data Mining (Khai phá dữ liệu): Các kỹ thuật dùng để khai phá dữ liệu và phát hiện tri thức như phân loại (Classification), phân nhóm (Clustering), phát hiện luật kết hợp (Association Rule), Dự đoán (Prediction),...
- + Business Analyst (Phân tích kinh Doanh): Các nhà lãnh đạo Doanh nghiệp đưa ra những quyết định chiến lược đối với hoạt động kinh doanh của doanh nghiệp.

### 3. Quy trình của Business Intelligence

Các bước trong quy trình kinh doanh thông minh gồm có 6 bước như sau:



- + Data Sources: Trong tầng đầu tiên của thành phần kiến trúc hệ thống BI, cần phải tập hợp và tích hợp các dữ liệu được chứa trong nhiều nguồn trực tiếp và nguồn gián tiếp không đồng nhất về xuất xứ và loại.
- + Data Warehouse: Chỗ chứa trước tiên nhất cho việc phát triển kiến trúc của hệ BI. Khối dữ liệu là các hệ thống thu thập tất cả các dữ liệu yêu cầu bởi một phòng ban nào đó của công ty như tiếp thị, đánh giá, cho mục đích phân tích một vài chức năng của hệ thống BI.
- + Data Mining: Đây là phần rất quan trọng trong hệ thống BI, là các phần sẽ biến đổi từ dữ liệu thô, khai thác những thông tin cần thiết để đưa ra và hỗ trợ trong việc ra quyết định. Bao gồm các kỹ thuật trích xuất thông tin, tri thức từ tập dữ liệu, gồm cả

các mô hình toán học cho việc nhận dạng mẫu, học máy và các kỹ thuật của khai phá dữ liệu.

- + Data Analysis Reporting: Thành phần tối ưu hóa cho phép xác định giải pháp tốt nhất từ tập hợp các hành động liên quan. Tập các hành động này có thể rất rộng và đôi khi không xác định. Từ đó tạo nên các báo cáo phân tích theo các phía cạnh.
- + Decision: Việc lựa chọn và thực thi phương thức quyết định nào đó dựa trên sự tính toán, so sánh đối chiếu của các phương thức toán học. Tuy nhiên, mặc dù cách thức lựa chọn được thông qua do cách thức toán học, việc quyết định theo hướng nào đó lại phụ thuộc vào người ra quyết định.

#### 4. Lợi ích của Business Intelligence

BI giúp cho các doanh nghiệp sử dụng thông tin một cách hiệu quả, chính xác để thích ứng với môi trường thay đổi liên tục và cạnh tranh khốc liệt trong kinh doanh:

- + Hỗ trợ nhà quản trị tối đa trong việc đưa ra các quyết định kinh doanh nhanh chóng, kịp thời, hiệu quả.
- + Xác định được vị thế và khả năng cạnh tranh của doanh nghiệp.
- + Phân tích hành vi khách hàng, Xác định mục đích và chiến lược Marketing.
- + Dự đoán tương lai của doanh nghiệp và xây dựng chiến lược kinh doanh.
- + Giữ chân được khách hàng cũ và dự đoán khách hàng tiềm năng.
- + Đáp ứng nhu cầu thu thập báo cáo của các bộ phận và cung cấp cái nhìn tổng thể toàn doanh nghiệp.
- + Hỗ trợ tối đa công tác điều hành, tiết kiệm thời gian và chi phí quản trị.
- + Góp phần thay đổi kỹ năng điều hành, phục vụ khách hàng tốt hơn.
- + Hỗ trợ người dùng nội bộ trong đánh giá, cải thiện và tối ưu hóa khả năng cũng như quy trình hoạt động của tổ chức.

Một số công cụ BI hiện hành:

- + Power BI
- + Tableau
- + Datapine
- + Sisense
- + Yellowfin BI
- + ...

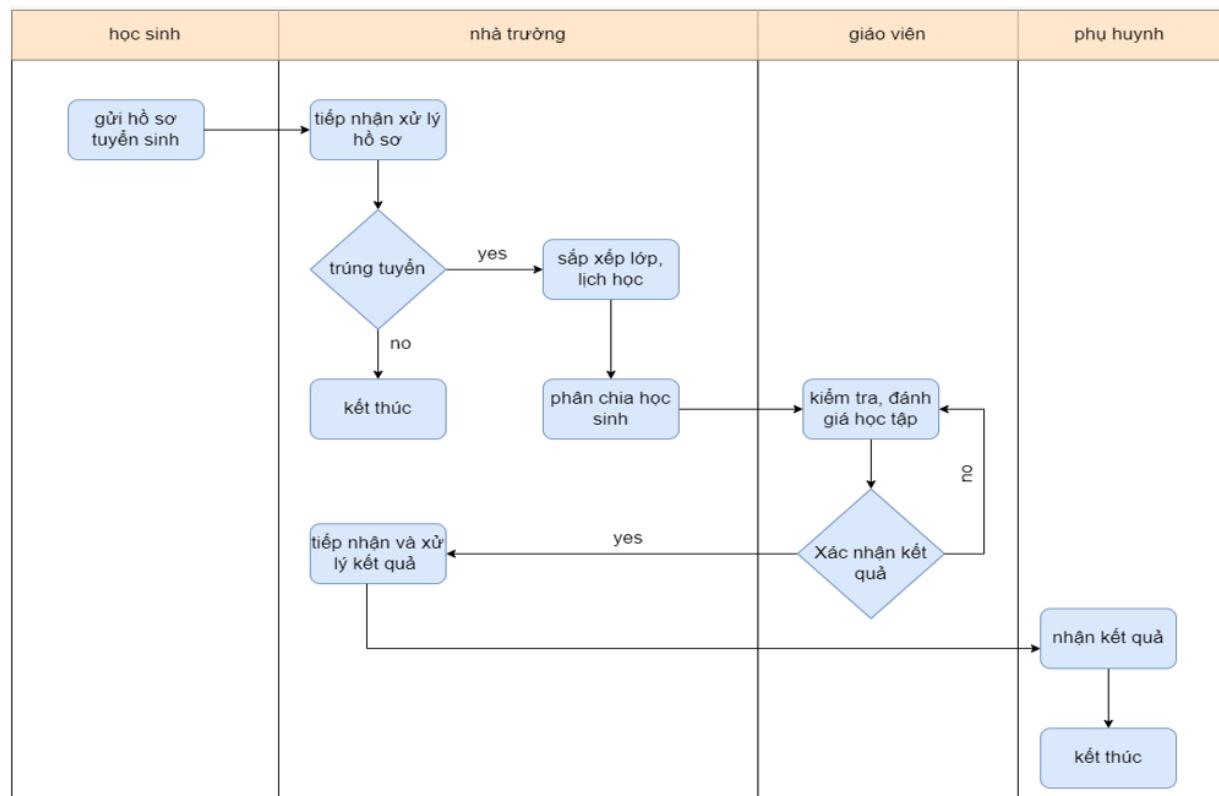
## CHƯƠNG 3: ỨNG DỤNG DATA WAREHOUSE VÀ BI VÀO GIÁO DỤC

### I. Khảo sát

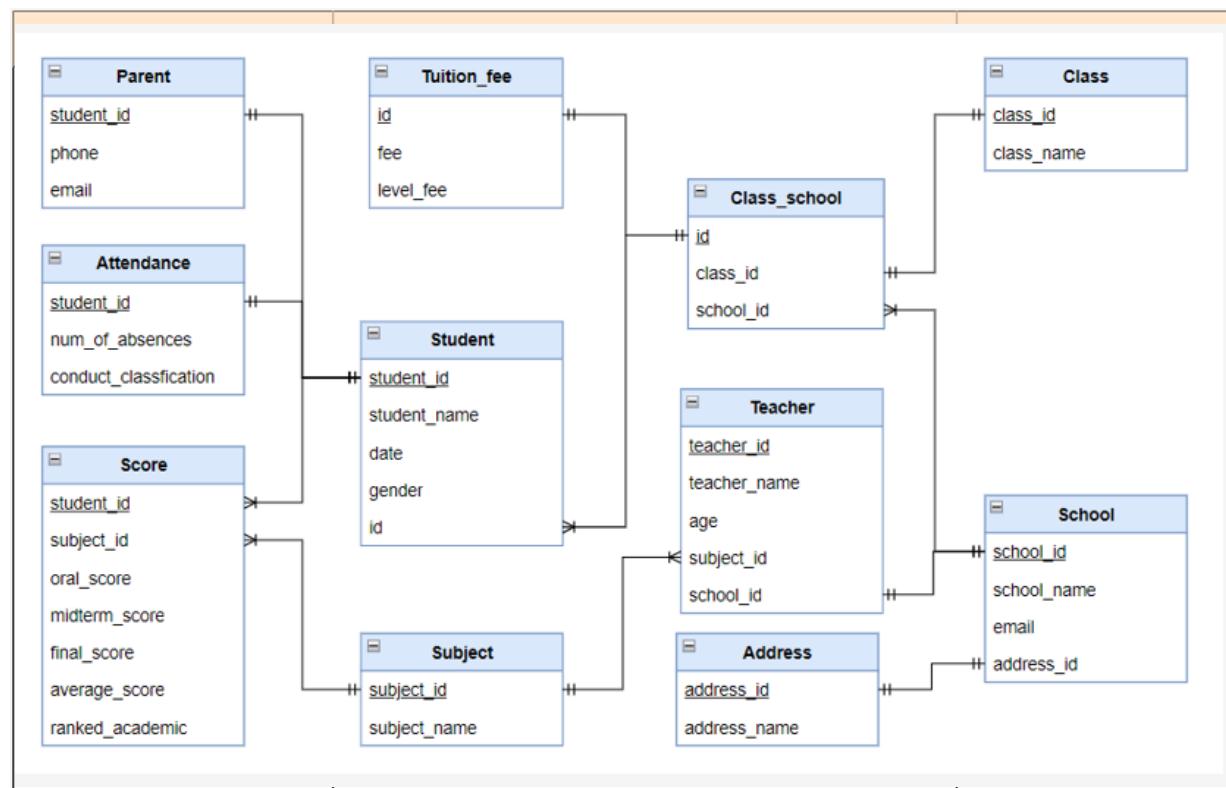
#### 1. Tổng quan

- Quản lý thông tin trong lĩnh vực giáo dục là quá trình thu thập, lưu trữ, tổ chức và xử lý thông tin liên quan đến hệ thống giáo dục.
- Nó bao gồm việc quản lý thông tin về học sinh, sinh viên, giáo viên, chương trình học, học phí và đánh giá trường học, giáo viên, học sinh trong lĩnh vực giáo dục, ...
- Tại sao nên sử dụng ?
  - Tăng cường khả năng ra quyết định

- Nâng cao hiệu suất và chất lượng giáo dục
  - Tối ưu hóa tài nguyên
  - Chức năng
    - Thu thập thông tin
    - Lưu trữ và tổ chức thông tin
    - Xử lý và phân tích thông tin
    - Bảo mật thông tin
    - Chia sẻ thông tin
  - Thực trạng : trên thực tế, nhiều trường học và cơ sở giáo dục ở Việt Nam đã bắt đầu sử dụng hệ thống quản lý thông tin trong giáo dục. Tuy nhiên, việc triển khai và sử dụng quản lý thông tin giáo dục vẫn chưa đồng nhất và phổ biến.
2. Quy trình nghiệp vụ
- Quy trình xử lý học tập :



- Quy trình xử lý học phí :
- ER Diagram :



### 3. Quy mô dữ liệu

- Về chi tiết các bảng dữ liệu :
- Các file dữ liệu là thông tin dữ liệu giáo dục tại các trường học tại thành phố

File	Chi tiết	Mô tả
address	30 rows , 2 columns, 10KB	thông tin địa chỉ của các trường học
attendance	300.000 rows, 2 columns , 4,085KB	thông tin số buổi nghỉ học của từng học sinh
class_school	7560 rows, 3columns , 141KB	thông tin mã trường học và mã lớp học tương ứng
classes	90 rows , 2 columns, 10KB	thông tin tên lớp học
parents	300.000 rows, 3 columns, 11,584 KB	thông tin phụ huynh học sinh và thông tin liên lạc
schools	252 rows, 4columns, 22KB	thông tin trường học
students	300.000 rows, 5 columns , 10,325KB	thông tin học sinh
subjects	12 rows, 2 columns, 9KB	các môn học
teachers	5653 rows, 5 columns, 227KB	thông tin giáo viên
Scores	3,600,000 rows, 6 columns , 140,967KB	điểm các môn học của từng học sinh
Tuition_fee	7560 rows, 2 columns , 198KB	thông tin học phí từng lớp học

Hà Nội

- Kích thước : 163 MB

### 4. Yêu cầu phân tích

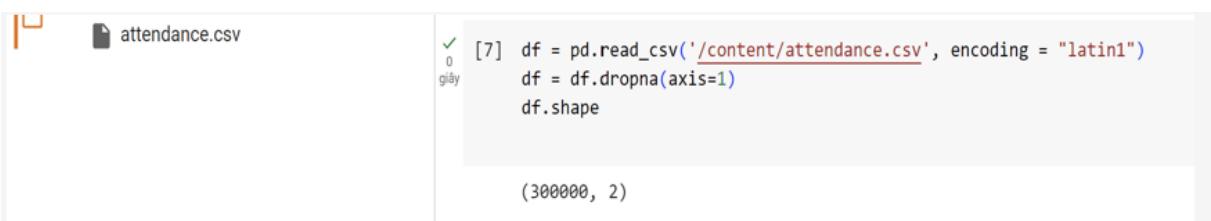
Chúng em chọn phân tích bộ dữ liệu theo 3 mặt sau

- Scores :
  - Phân tích điểm số trung bình
  - Điểm số theo môn học
  - Phân loại chất lượng học sinh
- Attendance :
  - Phân tích số buổi vắng mặt của học sinh
  - Phân tích theo lớp học , trường học
  - Phân tích tương quan giữa số buổi nghỉ học và điểm số
- Tuition fee :
  - Phân tích tổng quan
  - Học phí theo từng lớp học
  - Học phí theo trường học
  - Phân tích tương quan giữa học phí và điểm số

## II. Phân tích thiết kế

### 1. Data Exploration

- Bảng Attendance :



```
[7] df = pd.read_csv('/content/attendance.csv', encoding = "latin1")
df = df.dropna(axis=1)
df.shape

(300000, 2)
```

## Biểu

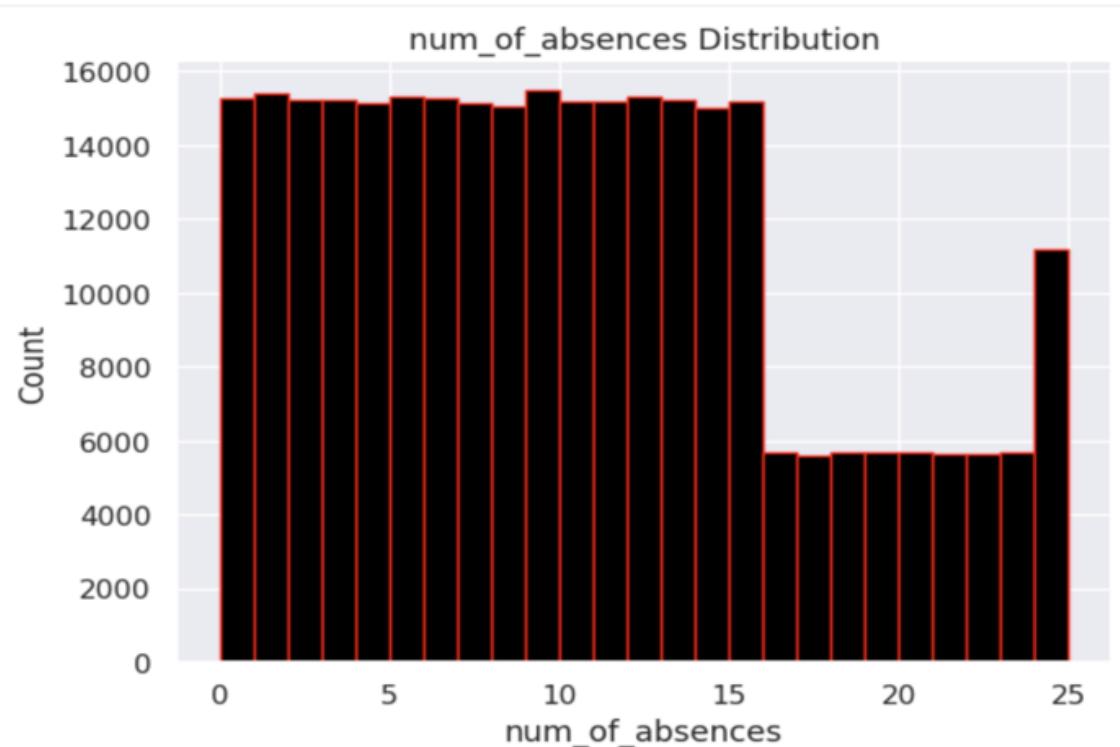
đò

[8] df.describe()

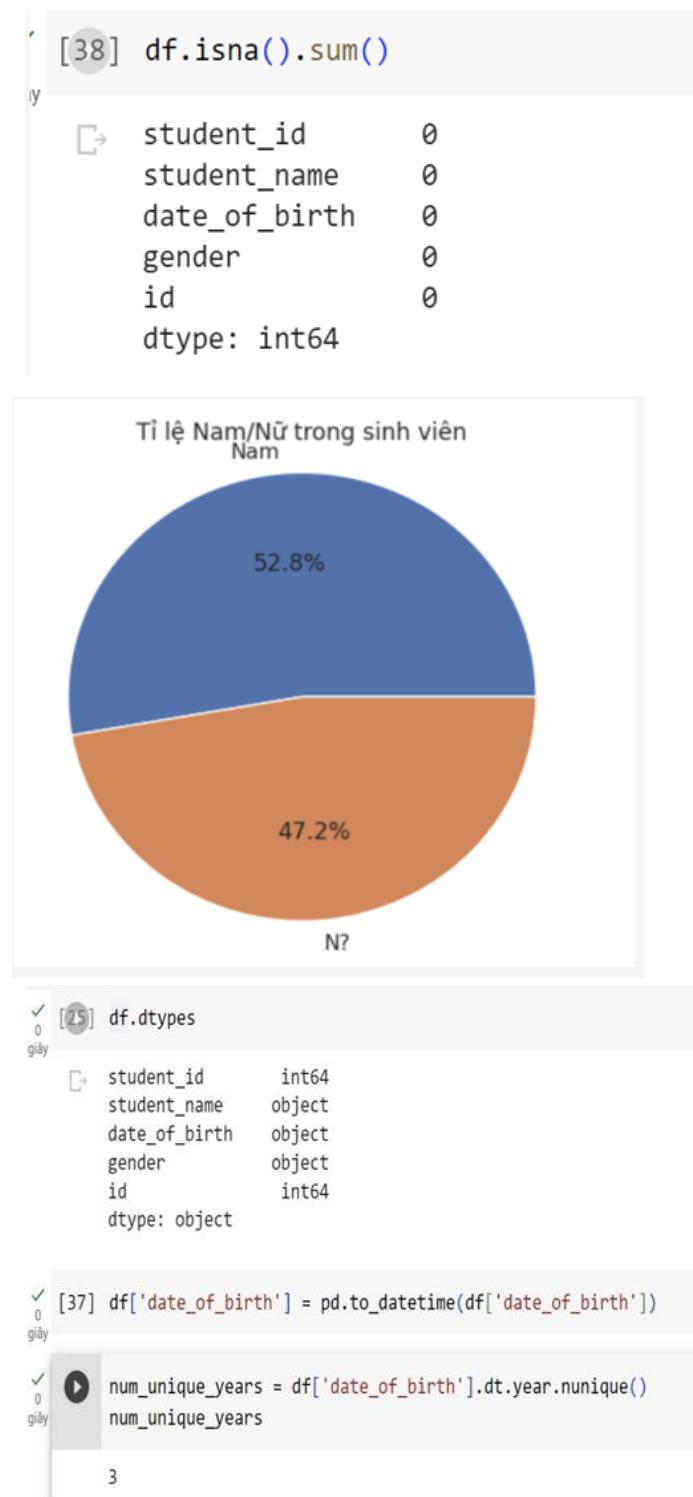


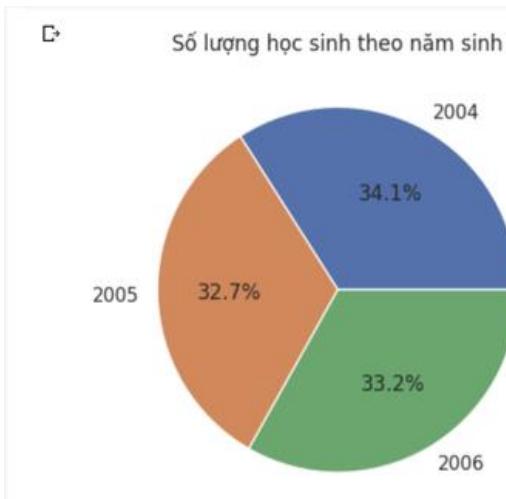
	student_id	num_of_absences
<b>count</b>	300000.000000	300000.000000
<b>mean</b>	150000.500000	9.933983
<b>std</b>	86602.684716	6.679954
<b>min</b>	1.000000	0.000000
<b>25%</b>	75000.750000	4.000000
<b>50%</b>	150000.500000	9.000000
<b>75%</b>	225000.250000	14.000000
<b>max</b>	300000.000000	25.000000

histogram thể hiện phân phối số lượng buổi nghỉ học :



## - Bảng Students





- Bảng Tuition Fee

	<b>id</b>	<b>fee</b>
0	1.0	820000.0
1	2.0	610000.0
2	3.0	610000.0
3	4.0	770000.0
4	5.0	910000.0
...	...	...
22585	NaN	NaN
22586	NaN	NaN
22587	NaN	NaN
22588	NaN	NaN
22589	NaN	NaN

22590 rows × 2 columns

Trước :

Xử lý :

```
[ ] data.isna().sum() #Đếm giá trị NaN ở các cột
```

```
id      15030
fee     15030
dtype: int64
```

	<b>id</b>	<b>fee</b>
0	1.0	820000.0
1	2.0	610000.0
2	3.0	610000.0
3	4.0	770000.0
4	5.0	910000.0
...	...	...
<b>7555</b>	7556.0	1000000.0
<b>7556</b>	7557.0	700000.0
<b>7557</b>	7558.0	830000.0
<b>7558</b>	7559.0	950000.0
<b>7559</b>	7560.0	730000.0

7560 rows × 2 columns

Sau :

⇒ Có 15030 dòng null và có 66,53% số dòng bị lỗi

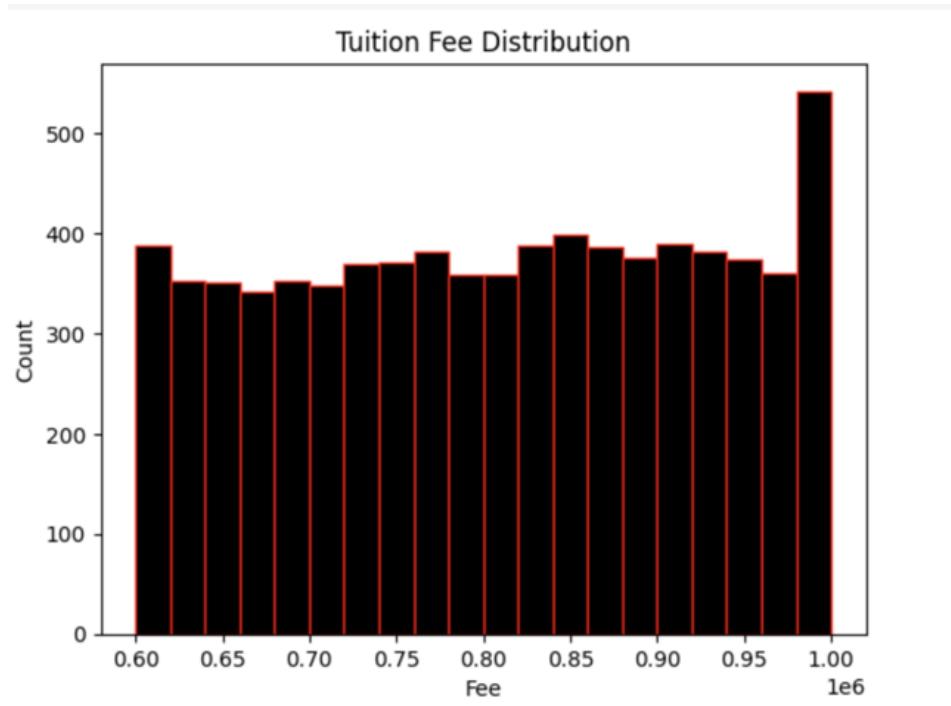
	<b>id</b>	<b>fee</b>
0	1	820000
1	2	610000
2	3	610000
3	4	770000
4	5	910000
...	...	...
<b>7555</b>	7556	1000000
<b>7556</b>	7557	700000
<b>7557</b>	7558	830000
<b>7558</b>	7559	950000
<b>7559</b>	7560	730000

7560 rows × 2 columns

```
[ ] df.describe()
```

	<b>id</b>	<b>fee</b>
<b>count</b>	7560.00000	7560.00000
<b>mean</b>	3780.50000	801798.941799
<b>std</b>	2182.52835	117678.145117
<b>min</b>	1.00000	600000.000000
<b>25%</b>	1890.75000	700000.000000
<b>50%</b>	3780.50000	800000.000000
<b>75%</b>	5670.25000	900000.000000
<b>max</b>	7560.00000	1000000.000000

Phân phối dữ liệu học phí bằng biểu đồ histogram :



- Bảng teachers

```
[49] te.isna().sum()
```

```
teacher_id      0
teacher_name    0
age             0
subject_id     0
school_id      0
dtype: int64
```

```
[50] te.dtypes
```

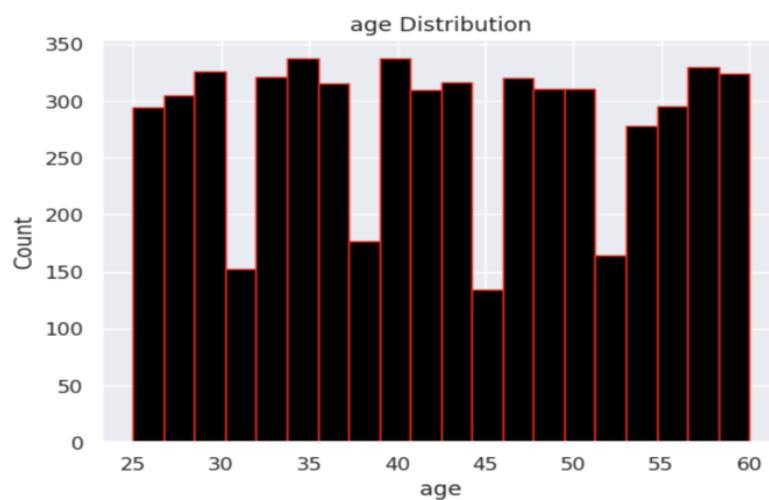
```
teacher_id      int64
teacher_name    object
age             int64
subject_id     int64
school_id      int64
dtype: object
```

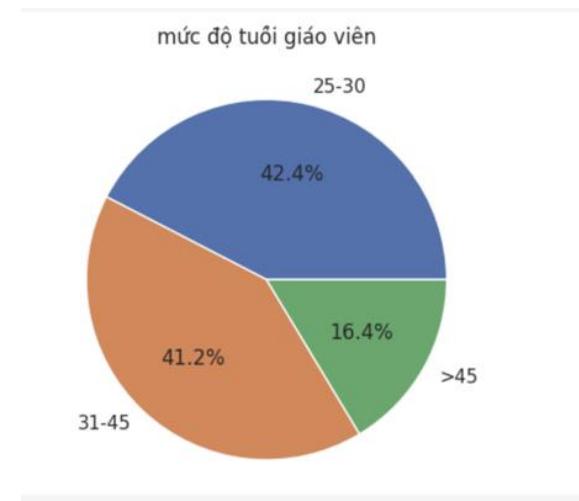
```
✓ [45] te['age'].describe()
```

giây

```
count    5653.000000
mean     42.452326
std      10.355520
min      25.000000
25%     34.000000
50%     42.000000
75%     51.000000
max      60.000000
Name: age, dtype: float64
```

Histogram phân phối tuổi của giáo viên :





- Bảng Scores

	oral_score	midterm_score	final_score
count	300000.000000	300000.000000	300000.000000
mean	5.700660	4.996913	7.708427
std	2.434579	2.582527	1.178266
min	1.000000	1.000000	6.000000
25%	4.000000	3.000000	6.000000
50%	6.000000	5.000000	8.000000
75%	8.000000	7.000000	9.000000
max	9.000000	9.000000	9.000000

Phân tích tương

quan :

```
▶ df1 = df[['oral_score','midterm_score','final_score']].tail(300000)
coleration_matrix = df1.corr()
coleration_matrix
```

	oral_score	midterm_score	final_score
oral_score	1.000000	0.002990	-0.198236
midterm_score	0.002990	1.000000	-0.001244
final_score	-0.198236	-0.001244	1.000000

Phân bố  
môn học

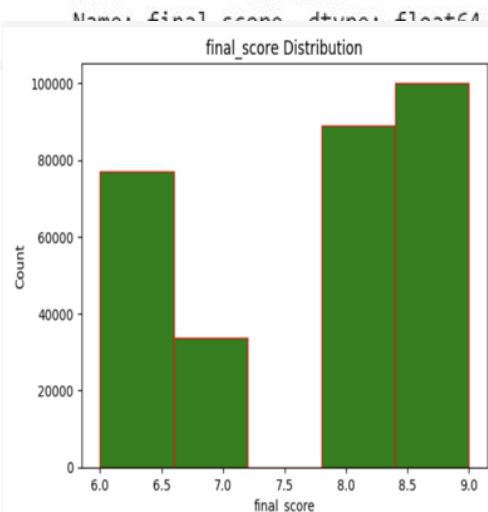
điểm cuối kì của  
ID=3

```
[ ]  
# Merge hai DataFrame theo cột 'student_id'  
merged_scores = pd.merge(df1, df2, on='student_id')  
  
# Tính ma trận hệ số tương quan  
correlation_matrix = merged_scores.corr()  
  
# In ra ma trận hệ số tương quan  
print(correlation_matrix)
```

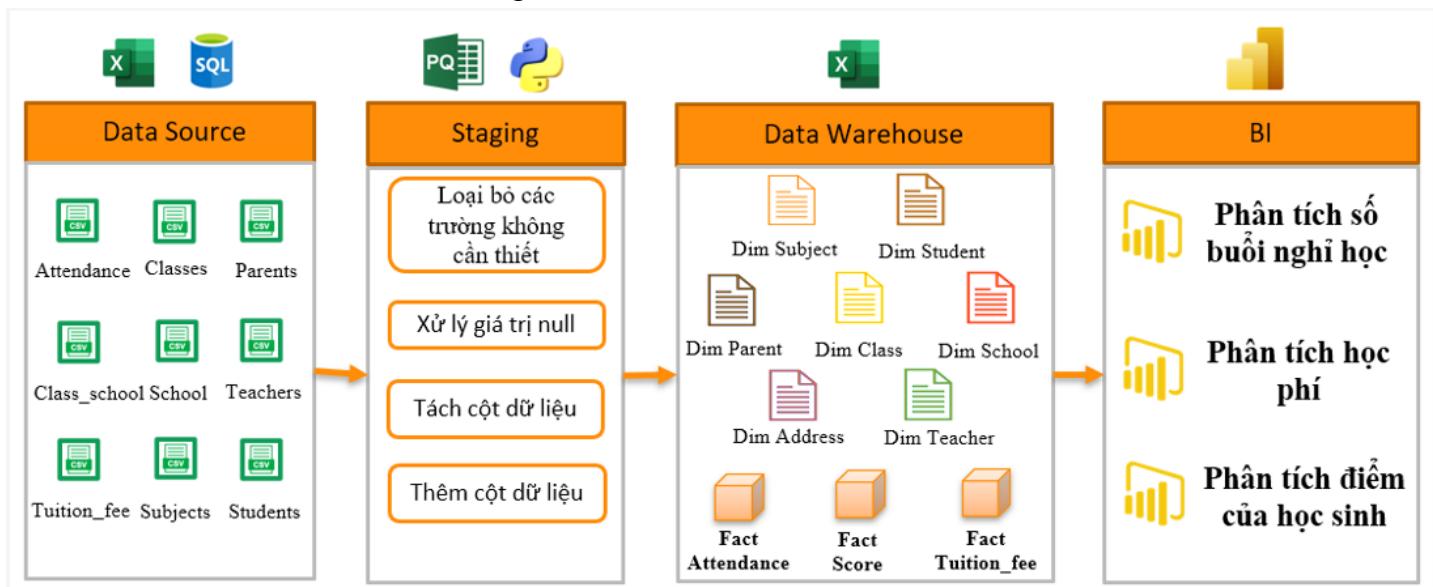
	student_id	final_score_x	final_score_y
student_id	1.000000	0.005293	-0.000966
final_score_x	0.005293	1.000000	0.001814
final_score_y	-0.000966	0.001814	1.000000

```
[28] df_final = df.tail(300000)  
df3 = df_final['final_score']  
df3.describe()
```

count	300000.000000
mean	7.708427
std	1.178266
min	6.000000
25%	6.000000
50%	8.000000
75%	9.000000
max	9.000000



2. Kiến trúc Data Warehouse
3. Các hoạt động ETL



- Tóm tắt các hoạt động ETL dữ liệu
  - Trích xuất dữ liệu từ các file Excel
  - Làm sạch dữ liệu
  - Loại bỏ các trường không cần thiết
  - Định dạng lại kiểu dữ liệu
  - Thêm cột dữ liệu
  - Load dữ liệu vào SQL server
- Bảng Students ( định dạng lại kiểu dữ liệu )
  - Trước

```
[ ] df.dtypes
student_id      int64
student_name    object
date_of_birth   object
gender          object
id              int64
dtype: object

[ ] df['date_of_birth'] = pd.to_datetime(df['date_of_birth'])
```

- Sau

```
[8] df.dtypes
y
student_id           int64
student_name          object
date_of_birth        datetime64[ns]
gender                object
id                   int64
dtype: object
```

- Bảng address ( xóa các trường dữ liệu không cần thiết)

- Trước

address_id	address_name	Column3	Column4	Column5	Column6
1	Ba Đình	null	null	null	null
2	Đống Đa	null	null	null	null
3	Hai Bà Trưng	null	null	null	null
4	Hoàn Kiếm	null	null	null	null
5	Tây Hồ	null	null	null	null
6	Thanh Xuân	null	null	null	null
7	Hoàng Mai	null	null	null	null
8	Long Biên	null	null	null	null
9	Cầu Giấy	null	null	null	null
10	Bắc Từ Liêm	null	null	null	null
11	Nam Từ Liêm	null	null	null	null
12	Hà Đông	null	null	null	null
13	Thanh Trì	null	null	null	null
14	Đông Anh	null	null	null	null
15	Thường Tín	null	null	null	null
16	Mê Linh	null	null	null	null

- Sau

address_id	address_name
1	Ba Đình
2	Đống Đa
3	Hai Bà Trưng
4	Hoàn Kiếm
5	Tây Hồ
6	Thanh Xuân
7	Hoàng Mai
8	Long Biên
9	Cầu Giấy
10	Bắc Từ Liêm
11	Nam Từ Liêm
12	Hà Đông
13	Thanh Trì
14	Đông Anh

- Bảng school ( xóa các trường dữ liệu không cần thiết )

- Trước
- Sau

1 <sup>2</sup> 3 school_id	A <sup>2</sup> C school_name	A <sup>2</sup> C email	1 <sup>2</sup> 3 address_id
1	Trường THPT BA VÌ	http://thptbaivi.edu.vn	3
2	Trường THPT ĐOÀN KẾT - HAI BÀ TRƯNG	http://thptdonket-habstrung.edu.vn	3
3	Trường THPT Nguyễn Văn Cừ	http://thptnguyenvancu.edu.vn	3
4	Trường THPT Lê Quý Đôn - Đồng Đa	http://thptlequydon-dongda.edu.vn	3
5	Trường THPT Hoàng Cầu	http://thpthoangcau.edu.vn	3
6	Trường THPT Xuân Mai	http://thptxuonmai.edu.vn	3
7	Trường THPT Thanh Oai A	http://thptthanhhoa.edu.vn	3
8	Trường THPT Nguyễn Du Thanh Oai	http://thptnguyenduthanhhoai.edu.vn	3
9	Trường THPT- THCS Hà Thành	http://thpt-thcshtt.edu.vn	1
10	Trường THPT Đinh Tiên Hoàng - Ba Đình	http://thptdinhthieng-badinh.edu.vn	1
11	Trường THPT Hồ Tùng Mậu	http://thpthotungmau.edu.vn	1
12	Trường THPT Nguyễn Trãi - Ba Đình	http://thptnguyentroi-badinh.edu.vn	1
13	Trường THPT Phạm Hồng Thái	http://thptphamhongthoi.edu.vn	1

1 <sup>2</sup> 3 school_id	A <sup>2</sup> C school_name	1 <sup>2</sup> 3 address_id
1	Trường THPT BA VÌ	3
2	Trường THPT ĐOÀN KẾT - HAI BÀ TRƯNG	3
3	Trường THPT Nguyễn Văn Cừ	3
4	Trường THPT Lê Quý Đôn - Đồng Đa	3
5	Trường THPT Hoàng Cầu	3
6	Trường THPT Xuân Mai	3
7	Trường THPT Thanh Oai A	3
8	Trường THPT Nguyễn Du Thanh Oai	3
9	Trường THPT- THCS Hà Thành	1
10	Trường THPT Đinh Tiên Hoàng - Ba Đình	1
11	Trường THPT Hồ Tùng Mậu	1
12	Trường THPT Nguyễn Trãi - Ba Đình	1
13	Trường THPT Phạm Hồng Thái	1
14	Trường THPT Phan Đình Phùng	1
15	Trường THPT Thực Nghiệm	1

-Bảng  
Students (

xóa các trường dữ liệu không cần thiết )

- Trước

student_id	student_name	date_of_birth	gender	id
1	Đàm Quang Duy	1/1/2004	Nữ	5887
2	Tiết Bảo Ngọc	1/1/2004	Nữ	5438
3	Giang Vĩnh Nguyên	1/1/2004	Nữ	7171
4	Lê Dương Bảo Anh	1/1/2004	Nữ	6283
5	Trần Lê Gia Bảo	1/1/2004	Nữ	6152
6	Phạm Văn Huy	1/1/2004	Nam	6211
7	Trần Đỗ Diệp Anh	1/1/2004	Nam	5158
8	Đàm Văn Đức	1/1/2004	Nam	7080
9	Phạm Hải Nam	1/1/2004	Nam	6998
10	Hoàng Minh Bảo Hân	1/1/2004	Nữ	6689

- Sau

student_id	date_of_birth	gender	id
1	1/1/2004	Nữ	5887
2	1/1/2004	Nữ	5438
3	1/1/2004	Nữ	7171
4	1/1/2004	Nữ	6283
5	1/1/2004	Nữ	6152
6	1/1/2004	Nam	6211
7	1/1/2004	Nam	5158
8	1/1/2004	Nam	7080
9	1/1/2004	Nam	6998
10	1/1/2004	Nữ	6689

- Bảng Tuition Fee (thêm cột dữ liệu với add conditional columns)
  - Trước

id	fee
1	820000
2	610000
3	610000
4	770000
5	910000

- Sau

1 <sup>2</sup> 3 id	1 <sup>2</sup> 3 fee	ABC 1 <sup>2</sup> 3 muc_hoc_phi
1	1	820000 chung
2	2	610000 thấp
3	3	610000 thấp
4	4	770000 chung
5	5	910000 cao

- Bảng Scores ( thêm cột dữ liệu với custom columns)

- Trước
- Sau

student_id	1 <sup>2</sup> 3 subject_id	1 <sup>2</sup> 3 oral_score	1 <sup>2</sup> 3 midterm_score	1 <sup>2</sup> 3 final_score	
1	4	7	3	8	
2	4	5	8	4	
3	4	4	9	9	
4	4	5	6	3	
5	4	2	5	8	
6	4	4	5	9	

student_id	1 <sup>2</sup> 3 subject_id	1 <sup>2</sup> 3 oral_score	1 <sup>2</sup> 3 midterm_score	1 <sup>2</sup> 3 final_score	ABC 1 <sup>2</sup> 3 avg_score
1	4	7	3	8	6.1
2	4	5	8	4	5.55
3	4	4	9	9	8.25
4	4	5	6	3	4.35
5	4	2	5	8	6.05
6	4	4	5	9	6.85

Thêm cột dữ liệu với coditional columns

- Trước

student_id	1 <sup>2</sup> 3 subject_id	1 <sup>2</sup> 3 oral_score	1 <sup>2</sup> 3 midterm_score	1 <sup>2</sup> 3 final_score	
1	4	7	3	8	
2	4	5	8	4	
3	4	4	9	9	
4	4	5	6	3	
5	4	2	5	8	
6	4	4	5	9	

- Sau

student_id	subject_id	oral_score	midterm_score	final_score	avg_score	hoc_luc
377	5	7	7	1 4	Yếu	
900	5	7	7	1 4	Yếu	
2216	5	7	7	1 4	Yếu	
2417	5	7	7	1 4	Yếu	
2564	5	7	7	1 4	Yếu	
3218	5	7	7	1 4	Yếu	
3363	5	7	7	1 4	Yếu	
4364	5	7	7	1 4	Yếu	
6592	5	7	7	1 4	Yếu	
6840	5	7	7	1 4	Yếu	
7490	5	7	7	1 4	Yếu	
8028	5	7	7	1 4	Yếu	

- Bảng Attendance ( thêm cột dữ liệu với add coditional columns)

- Trước

1 <sup>2</sup> 3 student_id	1 <sup>2</sup> 3 num_of_absences
1	2
2	12
3	11
4	13
5	5
6	12
7	6
8	12

- Sau

1 <sup>2</sup> 3 student_id	1 <sup>2</sup> 3 num_of_absences	ABC 123 NumOfAbsences	ABC 123 Hanh_kiem
1	2	nghỉ rất ít hoặc không nghỉ	T
2	12	nghỉ nhiều (6-15)	K
3	11	nghỉ nhiều (6-15)	K
4	13	nghỉ nhiều (6-15)	K
5	5	nghỉ học ít (3-5)	T
6	12	nghỉ nhiều (6-15)	K
7	6	nghỉ nhiều (6-15)	K
8	12	nghỉ nhiều (6-15)	K
9	0	nghỉ rất ít hoặc không nghỉ	T
10	9	nghỉ nhiều (6-15)	K
11	15	nghỉ rất nhiều(>15)	TB
12	1	nghỉ rất ít hoặc không nghỉ	T

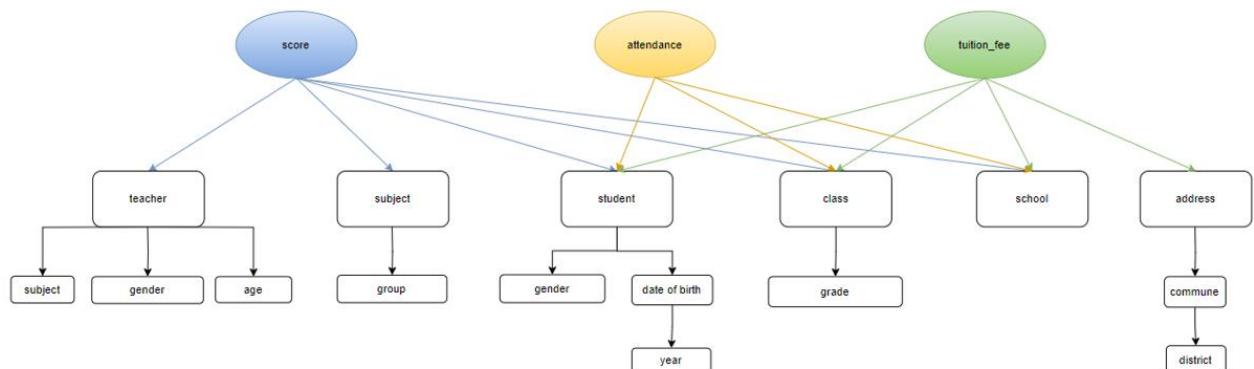
Hệ  
thống  
chiều khái niệm

address_id	address_name	30	class_name	90	school_id	252	Khối	3	subject_name	12
1	Ba Đình		10A1		1		10		Toán	
2	Đống Đa		10A2		2		11		Ngữ Văn	
3	Hai Bà Trưng		10A3		3		12		Tiếng Anh	
4	Hoàn Kiếm		10A4		4				Vật lí	
5	Tây Hồ		10A5		5				Hóa học	
6	Thanh Xuân		10A6		6				Sinh học	
7	Hoàng Mai		10A7		7				Lịch sử	
8	Long Biên		10A8		8				Địa lí	
9	Cầu Giấy		10A9		9				GDCD	
10	Bắc Từ Liêm		10A10		10				Tin học	
11	Nam Từ Liêm		10C1		11				GDQP-AN	
12	Hà Đông		10C2		12				Thể dục	
13	Thanh Trì		10C3		13					
14	Đông Anh		10C4		14					
15	Thường Tín		10C5		15					
16	Mê Linh		10C6		16					
17	Sóc Sơn		10C7		17					
18	Thạch Thất		10C8		18					
19	Quốc Oai		10C9		19					
20	Thanh Oai		10C10		20					
21	Đan Phượng		10D1		21					
22	Gia Lâm		10D2		22					
23	Chương Mỹ		10D3		23					
24	Hoài Đức				24					
25	Ba Vì				25					
26	Mỹ Đức				26					
27	Phú Xuyên				27					
28	Ứng Hòa				28					
29	Sơn Tây									
30	Phúc Thọ									

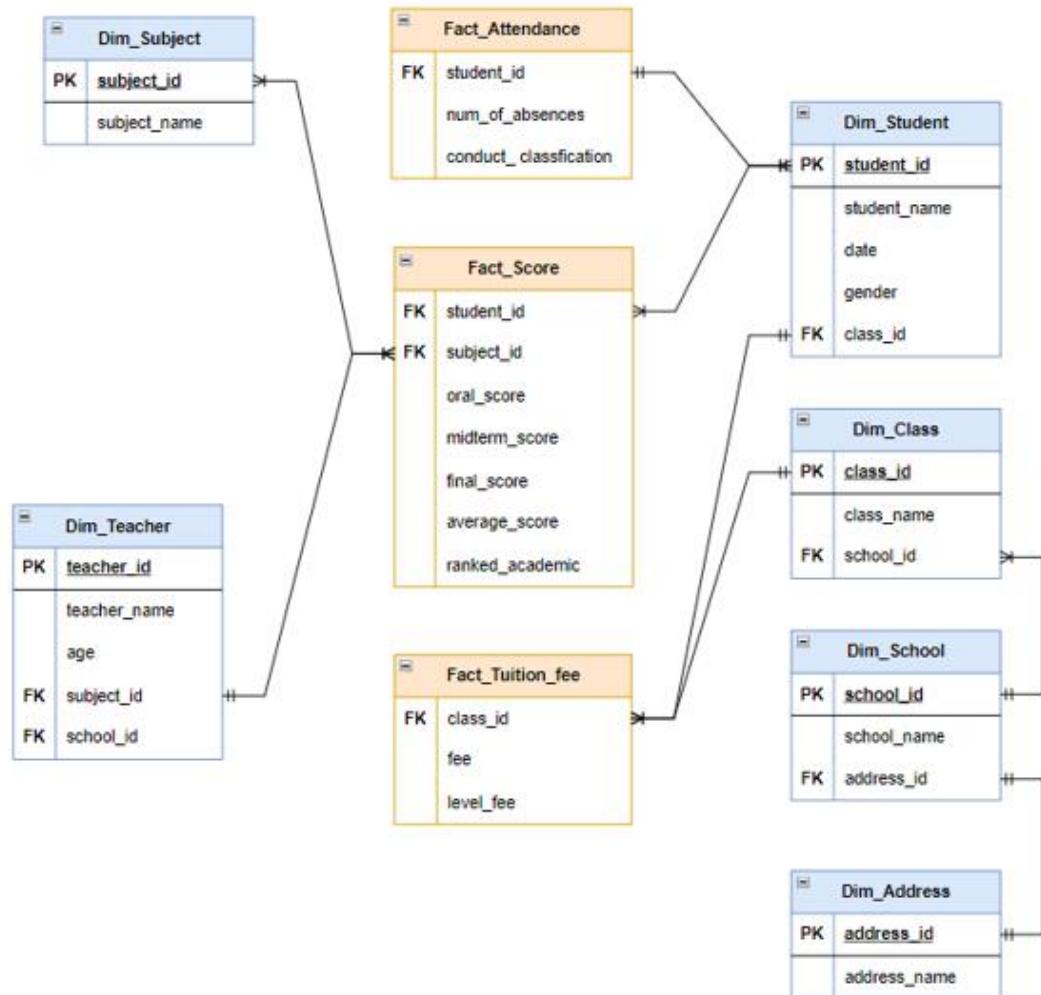
date_of_birth	3	gender	2	NumOfAbsences	4	Hanh_kiem	3	subject_name	12	hoc_luc	4	muc_hoc_phi	3
	2004	Nam		nghỉ nhiều (6-15)		K		Toán		Yếu		cao	
	2005	Nữ		nghỉ rất nhiều(>15)		TB		Ngữ Văn		TB		chung	
	2006			nghỉ học ít (3-5)		T		Tiếng Anh		Khá		thấp	
				nghỉ rất ít hoặc không nghỉ				Vật lí		Giỏi			
								Hóa học					
								Sinh học					
								Lịch sử					
								Địa lí					
								GDCD					
								Tin học					
								GDQP-AN					
								Thể dục					

#### 4. Xây dựng data model OLAP

- Data model logic



- Data model vật lý



### 5. Xây dựng cơ sở dữ liệu OLAP

- Xây dựng qua OLAP Views
  - Tạo view Address

```
-- Tạo view address
create view v_dim_address
as
select address_id, address_name
from address

-- Chạy view address
select * from v_dim_address
```

	address_id	address_name
1	1	Ba Đình
2	2	Đống Đa
3	3	Hai Bà Trưng
4	4	Hoàn Kiếm
5	5	Tây Hồ
6	6	Thanh Xuân
7	7	Hoàng Mai
8	8	Long Biên
9	9	Cầu Giấy
10	10	Bắc Từ Liêm
11	11	Nam Từ Liêm
12	12	Hà Đông
13	13	Thanh Trì
14	14	Đông Anh
15	15	Thường Tín
16	16	Mê Linh
17	17	Sóc Sơn
18	18	Thạch Thất
19	19	Quốc Oai
20	20	Thanh Oai
21	21	Đan Phượng
22	22	Gia Lâm

- Tạo view School

```
-- Tạo view school
create view v_dim_school
as
select school_id, school_name, address_id
from schools

-- chạy view school
select * from v_dim_school
```

	school_id	school_name	address_id
1	1	Trường THPT BA VI	3
2	2	Trường THPT ĐOÀN KẾT - HAI BÀ TRUNG	3
3	3	Trường THPT Nguyễn Văn Cừ	3
4	4	Trường THPT Lê Quý Đôn - Đống Đa	3
5	5	Trường THPT Hoàng Cầu	3
6	6	Trường THPT Xuân Mai	3
7	7	Trường THPT Thanh Oai A	3
8	8	Trường THPT Nguyễn Du Thanh Oai	3
9	9	Trường THPT- THCS Hà Thành	1
10	10	Trường THPT Đinh Tiên Hoàng - Ba Đình	1
11	11	Trường THPT Hồ Tùng Mậu	1
12	12	Trường THPT Nguyễn Trãi - Ba Đình	1
13	13	Trường THPT Phạm Hồng Thái	1
14	14	Trường THPT Phan Đình Phùng	1
15	15	Trường THPT Thực Nghiêm	1
16	16	Trường THPT Văn Lang	1
17	17	Trường THPT Dân tộc nội trú	25
18	18	Trường THPT Ba Vì	25
19	19	Trường THPT Bát Bạt	25
20	20	Trường THPT Lương Thế Vinh - Ba Vì	25

- Tạo view Class

```
-- Tạo view class
create view v_dim_class
as
select c1.id, c2.class_name, c1.school_id
from class_school c1
inner join classes c2
on c1.class_id = c2.class_id

-- chạy view class
select * from v_dim_class
```

	<u>id</u>	<u>class_name</u>	<u>school_id</u>
1	1	10A1	1
2	2	10A1	2
3	3	10A1	3
4	4	10A1	4
5	5	10A1	5
6	6	10A1	6
7	7	10A1	7
8	8	10A1	8
9	9	10A1	9
10	1...	10A1	10
11	1...	10A1	11
12	1...	10A1	12
13	1...	10A1	13
14	1...	10A1	14
15	1...	10A1	15
16	1...	10A1	16
17	1...	10A1	17
18	1...	10A1	18
19	1...	10A1	19
20	2...	10A1	20

- Tạo view Student

```
-- Tạo view student
create view v_dim_student
as
select student_id, student_name, date_of_birth, gender, id
from students

-- chạy view student
select * from v_dim_student
```

	student_id	student_name	date_of_birth	gender	id
1	3781	Nguyễn Công Bảo	2004-01-14 00:00:00.000	Nam	7397
2	3782	Vũ Ngọc Hiếu Ngân	2004-01-14 00:00:00.000	Nữ	4491
3	3783	Lê Ngọc Quyên	2004-01-14 00:00:00.000	Nam	2941
4	3784	Phạm Thu Ngân	2004-01-14 00:00:00.000	Nam	4385
5	3785	Nguyễn Phạm Tuyết Nhi	2004-01-14 00:00:00.000	Nữ	4886
6	3786	Nguyễn Minh Thuận	2004-01-14 00:00:00.000	Nam	7160
7	3787	Phạm Ngọc Châu Giang	2004-01-14 00:00:00.000	Nam	4620
8	3788	Nguyễn Nhật Thành	2004-01-14 00:00:00.000	Nữ	2999
9	3789	Phạm Tú Uyên	2004-01-14 00:00:00.000	Nữ	6959
10	3790	Lưu Nhật Minh	2004-01-14 00:00:00.000	Nữ	7293
11	3791	Phạm Thuý Dương	2004-01-14 00:00:00.000	Nam	1412
12	3792	Vũ Minh Lương	2004-01-14 00:00:00.000	Nữ	5995
13	3793	Trần Quỳnh Giang	2004-01-14 00:00:00.000	Nam	598
14	3794	Hoàng Khánh Giang	2004-01-14 00:00:00.000	Nữ	4915
15	3795	Mai Phương Uyên	2004-01-14 00:00:00.000	Nữ	5798
16	3796	Phạm Phương Bảo Ngọc	2004-01-14 00:00:00.000	Nữ	3131
17	3797	Trần Vũ Thảo Nguyên	2004-01-14 00:00:00.000	Nữ	7079
18	3798	Bùi Nhật Dương	2004-01-14 00:00:00.000	Nữ	4544
19	3799	Trần Khánh Nam	2004-01-14 00:00:00.000	Nam	5047
20	3800	Nguyễn Thái Dương	2004-01-14 00:00:00.000	Nam	7460

- Tạo view Teacher

```
-- Tạo view teacher
create view v_dim_teacher
as
select teacher_id, teacher_name, age, subject_id, school_id
from teachers

-- chạy view teacher
select * from v_dim_teacher
```

	teacher_id	teacher_name	age	subject_id	school_id
1	1	Nguyễn Phú Đồng	56	4	147
2	2	Huỳnh Văn Nhứt	44	4	207
3	3	Nguyễn Thị Minh Hằng	41	4	190
4	4	Trần Thị Thanh Hảo	35	4	241
5	5	Lương Thị Phương	30	4	26
6	6	Hồ Thị Quỳnh Giang	50	8	7
7	7	Đỗ Thị Ngọc Lan	53	8	194
8	8	Nguyễn Thị Thanh Vân	40	8	38
9	9	Tạ Thị Hiệp	40	8	45
10	10	Võ Thị Minh Thúy	30	12	178
11	11	Nguyễn Việt Sơn	25	12	77
12	12	Đỗ Thị Thanh Tiên	44	12	29
13	13	Trần Văn Quang	58	5	112
14	14	Bùi Thanh Huyền	59	5	223
15	15	Nguyễn Thị Xuân Hương	57	5	71
16	16	Đoàn Thị Đương	34	5	190
17	17	Nguyễn Thuý Nữ Hiệp	45	5	24
18	18	Lê Thị Thu Tâm	58	5	240
19	19	Phan Thị Diễm Quyên	46	5	72
20	20	Nguyễn Thị Mỹ Tinh	28	5	226

- Tạo thêm cột điểm trung bình và học lực
- Tạo view Scores

```
-- tạo cột average_score, ranked_academic
ALTER TABLE scores
ADD average_score FLOAT;
UPDATE scores
SET average_score = (oral_score * 0.3 + midterm_score * 0.7) / 2 + final_score / 2

UPDATE scores
SET average_score = ROUND(average_score, 1);

ALTER TABLE scores
ADD ranked_academic NVARCHAR(255);

UPDATE scores
SET ranked_academic =
CASE
    WHEN average_score >= 8 THEN 'Gioi'
    WHEN average_score >= 6.5 THEN 'Kha'
    WHEN average_score >= 5 THEN 'Trung binh'
    ELSE 'Yeu'
END;
```

- Tạo thêm cột hạnh kiểm

```
-- tạo cột conduct_classification
ALTER TABLE attendance
ADD conduct_classification NVARCHAR(255);

UPDATE attendance
SET conduct_classification =
CASE
    WHEN num_of_absences > 15 THEN 'Trung binh'
    WHEN num_of_absences > 5 THEN 'Kha'
    ELSE 'Tot'
END;
```



- Tạo view Attendance

```
-- tạo fact attendance
create view v_fact_attendance
as
select student_id, num_of_absences, conduct_classification
from attendance

-- chạy fact attendance
select * from v_fact_attendance
```

	student_id	num_of_absences	conduct_classification
1	970	7	Kha
2	971	8	Kha
3	972	6	Kha
4	973	12	Kha
5	974	2	Tot
6	975	6	Kha
7	976	15	Kha
8	977	5	Tot
9	978	13	Kha
10	979	14	Kha
11	980	6	Kha
12	981	8	Kha
13	982	10	Kha
14	983	11	Kha
15	984	5	Tot
16	985	11	Kha
17	986	1	Tot
18	987	0	Tot
19	988	11	Kha
20	989	4	Tot

- Tạo view Tuition Fee

```
-- tạo fact tuition_fee
create view v_fact_tuition_fee
as
select id, fee, level_fee
from tuition_fee

-- chạy fact tuition_fee
select * from v_fact_tuition_fee
```

	id	fee	level_fee
1	1	820000	Binh Thuong
2	2	610000	Thap
3	3	610000	Thap
4	4	770000	Binh Thuong
5	5	910000	Cao
6	6	810000	Binh Thuong
7	7	670000	Thap
8	8	860000	Cao
9	9	660000	Thap
10	1...	760000	Binh Thuong
11	1...	690000	Thap
12	1...	910000	Cao
13	1...	730000	Thap
14	1...	650000	Thap
15	1...	780000	Binh Thuong
16	1...	790000	Binh Thuong
17	1...	850000	Cao
18	1...	790000	Binh Thuong
19	1...	910000	Cao
20	2...	930000	Cao

- Xây dựng OLAP qua thiết kế bảng : tạo bảng và đổ dữ liệu
  - Dim address

```

-- tạo table dim_address
create table dim_address(
address_id float not null,
address_name nvarchar(255) not null,
constraint pk_dim_address primary key(address_id)
)
go

-- đổ dữ liệu vào dim_address
INSERT INTO dim_address (address_id, address_name)
SELECT address_id, address_name
FROM v_dim_address;
:
```

```
select * from dim_address
```

	address_id	address_name
1	1	Ba Đình
2	2	Đống Đa
3	3	Hai Bà Trưng
4	4	Hoàn Kiếm
5	5	Tây Hồ
6	6	Thanh Xuân
7	7	Hoàng Mai
8	8	Long Biên
9	9	Cầu Giấy
10	10	Bắc Từ Liêm
11	11	Nam Từ Liêm
12	12	Hà Đông
13	13	Thanh Trì
14	14	Đông Anh
15	15	Thường Tín
16	16	Mê Linh
17	17	Sóc Sơn
18	18	Thạch Thất
19	19	Quốc Oai
20	20	Thanh Oai

Kết quả

- Dim School

```
-- tạo table dim_school
create table dim_school(
    school_id float not null,
    school_name nvarchar(255) not null,
    address_id float not null,
    constraint pk_dim_school primary key(school_id),
    constraint fk_dim_school_dim_address foreign key(address_id) references dim_address(address_id)
)
go

-- đổ dữ liệu vào dim_school
INSERT INTO dim_school (school_id, school_name, address_id)
SELECT school_id, school_name, address_id
FROM v_dim_school;

select * from dim_school
```



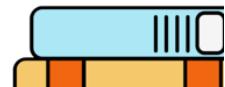
Kết quả :

	school_id	school_name	address_id
1	1	Trường THPT BA VĨ	3
2	2	Trường THPT ĐOÀN KẾT - HAI BÀ TRUNG	3
3	3	Trường THPT Nguyễn Văn Cừ	3
4	4	Trường THPT Lê Quý Đôn - Đống Đa	3
5	5	Trường THPT Hoàng Cầu	3
6	6	Trường THPT Xuân Mai	3
7	7	Trường THPT Thanh Oai A	3
8	8	Trường THPT Nguyễn Du Thanh Oai	3
9	9	Trường THPT- THCS Hà Thành	1
10	10	Trường THPT Đinh Tiên Hoàng - Ba Đình	1
11	11	Trường THPT Hồ Tùng Mậu	1
12	12	Trường THPT Nguyễn Trãi - Ba Đình	1
13	13	Trường THPT Phạm Hồng Thái	1
14	14	Trường THPT Phan Đình Phùng	1
15	15	Trường THPT Thực Nghiêm	1
16	16	Trường THPT Văn Lang	1
17	17	Trường THPT Dân tộc nội trú	25
18	18	Trường THPT Ba Vì	25
19	19	Trường THPT Bát Bạt	25
20	20	Trường THPT Lương Thế Vinh - Ba Vì	25

- Dim Class

```
-- tạo table dim_class
create table dim_class(
class_id float not null,
class_name nvarchar(255) not null,
school_id float not null,
constraint pk_dim_class primary key(class_id),
constraint fk_dim_class_dim_school foreign key(school_id) references dim_school(school_id)
)
go

-- đổ dữ liệu vào dim_class
INSERT INTO dim_class (class_id, class_name, school_id)
SELECT id, class_name, school_id
FROM v_dim_class;
select * from dim_class
```



Kết quả :

	class_id	class_name	school_id
1	1	10A1	1
2	2	10A1	2
3	3	10A1	3
4	4	10A1	4
5	5	10A1	5
6	6	10A1	6
7	7	10A1	7
8	8	10A1	8
9	9	10A1	9
10	10	10A1	10
11	11	10A1	11
12	12	10A1	12
13	13	10A1	13
14	14	10A1	14
15	15	10A1	15
16	16	10A1	16
17	17	10A1	17
18	18	10A1	18
19	19	10A1	19
20	20	10A1	20

- Dim Student

Kết quả :

```
-- tạo table dim_student
create table dim_student(
student_id float not null

```

	student_id	student_name	date_of_birth	gender	class_id
1	1	Đàm Quang Duy	2004-01-01 00:00:00.000	Nữ	5887
2	2	Tiết Bảo Ngọc	2004-01-01 00:00:00.000	Nữ	5438
3	3	Giang Vĩnh Nguyên	2004-01-01 00:00:00.000	Nữ	7171
4	4	Lê Dương Bảo Anh	2004-01-01 00:00:00.000	Nữ	6283
5	5	Trần Lê Gia Bảo	2004-01-01 00:00:00.000	Nữ	6152
6	6	Phạm Văn Huy	2004-01-01 00:00:00.000	Nam	6211
7	7	Trần Đỗ Diệp Anh	2004-01-01 00:00:00.000	Nam	5158
8	8	Đàm Văn Đức	2004-01-01 00:00:00.000	Nam	7080
9	9	Phạm Hải Nam	2004-01-01 00:00:00.000	Nam	6998
10	10	Hoàng Minh Bảo Hân	2004-01-01 00:00:00.000	Nữ	6689
11	11	Nguyễn Ngọc Khánh	2004-01-01 00:00:00.000	Nam	5239
12	12	Trần Hữu Anh Hào	2004-01-01 00:00:00.000	Nam	7282
13	13	Lê Minh Trí	2004-01-01 00:00:00.000	Nữ	5989
14	14	Trường Anh Kiệt	2004-01-01 00:00:00.000	Nam	5937
15	15	Phạm Phương Bảo Ngọc	2004-01-01 00:00:00.000	Nữ	5811
16	16	Nguyễn Đức Gia Hò	2004-01-01 00:00:00.000	Nam	6861
17	17	Hoàng Bảo Trung	2004-01-01 00:00:00.000	Nam	6345
18	18	Phạm Phương Linh	2004-01-01 00:00:00.000	Nam	6233
19	19	Phạm Ngọc Minh	2004-01-01 00:00:00.000	Nam	5309
20	20	Trần Thị Minh Khánh	2004-01-01 00:00:00.000	Nữ	6022

- Dim Subject

```
-- tạo table dim_subject
|create table dim_subject(
  subject_id float not null,
  subject_name nvarchar(255) not null,
  constraint pk_dim_subject primary key(subject_id)
)
go

-- đổ dữ liệu vào dim_subject
|INSERT INTO dim_subject (subject_id, subject_name)
  SELECT subject_id, subject_name
  FROM v_dim_subject;

select * from dim_subject
```



Kết quả:

	subject_id	subject_name
1	1	Toán
2	2	Ngữ Văn
3	3	Tiếng Anh
4	4	Vật lí
5	5	Hóa học
6	6	Sinh học
7	7	Lịch sử
8	8	Địa lí
9	9	GDCD
10	10	Tin học
11	11	GDQP-AN
12	12	Thể dục

- Dim teacher

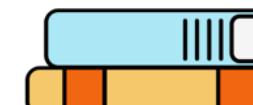
```
-- tạo table dim_teacher
create table dim_teacher(
teacher_id float not null,
teacher_name nvarchar(255) not null,
age float not null,
subject_id float not null,
school_id float not null,
constraint pk_dim_teacher primary key(teacher_id),
constraint fk_dim_teacher_dim_subject foreign key(subject_id) references dim_subject(subject_id)
constraint fk_dim_teacher_dim_school foreign key(school_id) references dim_school(school_id)
)
go

-- đổ dữ liệu vào dim_teacher
INSERT INTO dim_teacher (teacher_id, teacher_name, age, subject_id, school_id)
SELECT teacher_id, teacher_name, age, subject_id, school_id
FROM v_dim_teacher;

select * from dim_teacher

```

Kết quả :



	teacher_id	teacher_name	age	subject_id	school_id
1	1	Nguyễn Phú Đồng	56	4	147
2	2	Huỳnh Văn Nhứt	44	4	207
3	3	Nguyễn Thị Minh Hằng	41	4	190
4	4	Trần Thị Thanh Hào	35	4	241
5	5	Lương Thị Phương	30	4	26
6	6	Hồ Thị Quỳnh Giang	50	8	7
7	7	Đỗ Thị Ngọc Lan	53	8	194
8	8	Nguyễn Thị Thanh Vân	40	8	38
9	9	Tạ Thị Hiệp	40	8	45
10	10	Võ Thị Minh Thùy	30	12	178
11	11	Nguyễn Viết Sơn	25	12	77
12	12	Đỗ Thị Thanh Tiễn	44	12	29
13	13	Trần Văn Quang	58	5	112
14	14	Bùi Thanh Huyền	59	5	223
15	15	Nguyễn Thị Xuân Hư...	57	5	71
16	16	Đoàn Thị Đương	34	5	190
17	17	Nguyễn Thuỳ Nữ Hiệp	45	5	24
18	18	Lê Thị Thu Tâm	58	5	240
19	19	Phan Thị Diễm Quyên	46	5	72
20	20	Nguyễn Thị Mỹ Tính	28	5	226

- Fact Scores

```
-- tạo table fact_score
CREATE TABLE fact_score(
    student_id float not null,
    subject_id float not null,
    oral_score float not null,
    midterm_score float not null,
    final_score float not null,
    average_score float not null,
    ranked_academic nvarchar(255) not null,
    constraint fk_fact_score_dim_student foreign key(student_id) references dim_student(student_id)
    constraint fk_fact_score_dim_subject foreign key(subject_id) references dim_subject(subject_id)
)
go

-- đổ dữ liệu vào fact_score
<INSERT INTO fact_score (student_id, subject_id, oral_score, midterm_score, final_score, average_score, ranked_academic)
SELECT student_id, subject_id, oral_score, midterm_score, final_score, average_score, ranked_academic
FROM v_fact_score;

select * from fact_score
```



Kết quả :

- Fact Attendance

```
-- tạo table fact_attendance
create table fact_attendance(
student_id float not null,
num_of_absences float not null,
conduct_classification nvarchar(255) not null,
constraint fk_fact_attendance_dim_student foreign key(student_id) references dim_student(student_id)
)
go
```

```
-- đổ dữ liệu vào fact_attendance
INSERT INTO fact_attendance (student_id, num_of_absences, conduct_classification)
SELECT student_id, num_of_absences, conduct_classification
```

```
1 FROM v_fact_attendance;
2
3 select * from fact_attendance
```

4	1	2	8	3	6	5.3	Trung bình
5	1	7	5	8	3	5	Trung bình
6	1	9	2	3	4	3.3	Yếu
7	1	6	3	5	3	3.7	Yếu
8	1	1	4	4	4	4	Yếu
9	1	8	2	6	8	6.4	Trung bình
10	1	4	7	3	8	6.1	Trung bình
11	1	11	7	8	6	6.8	Kha
12	1	4	7	3	8	6.1	Trung bình
13	1	10	9	6	9	7.9	Kha
14	1	3	7	3	7	5.6	Trung bình
15	2	2	6	9	3	5.5	Trung bình
16	2	6	4	1	2	1.9	Yếu
17	2	6	4	1	2	1.9	Yếu
18	2	12	8	7	7	7.1	Kha
19	2	7	7	8	9	8.3	Gioi
20	2	3	5	5	3	4	Yếu

Kết quả :

	student_id	num_of_absences	conduct_classification
1	1	2	Tot
2	2	12	Kha
3	3	11	Kha
4	4	13	Kha
5	5	5	Tot
6	6	12	Kha
7	7	6	Kha
8	8	12	Kha
9	9	0	Tot
10	10	9	Kha
11	11	15	Kha
12	12	1	Tot
13	13	7	Kha
14	14	3	Tot
15	15	5	Tot
16	16	1	Tot
17	17	11	Kha
18	18	9	Kha
19	19	1	Tot
20	20	8	Kha

- Fact Tuition Fee

```
-- tạo table fact_tuition_fee
CREATE TABLE fact_tuition_fee(
    class_id float not null,
    fee float not null,
    level_fee nvarchar(255) not null,
    constraint fk_fact_score_dim_class foreign key(class_id) references dim_class(class_id)
)
go

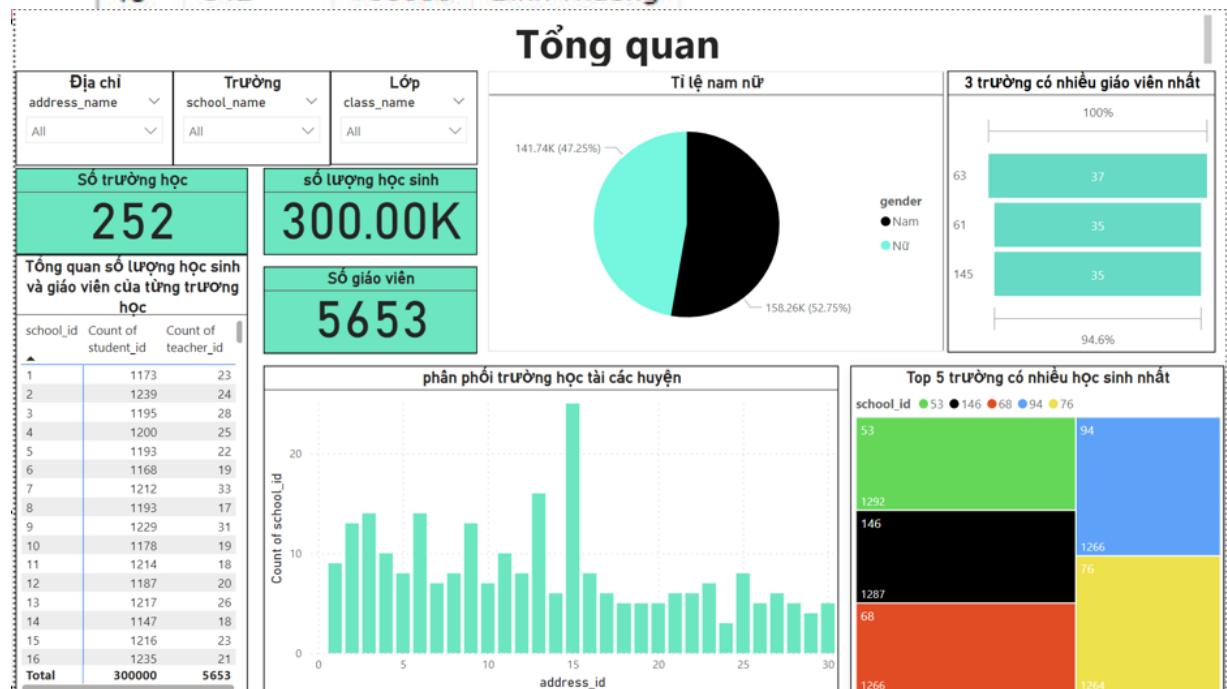
-- đổ dữ liệu vào fact_tuition_fee
INSERT INTO fact_tuition_fee (class_id, fee, level_fee)
SELECT id, fee, level_fee
FROM v_fact_tuition_fee;

select * from v_fact_tuition_fee
```



Kết quả :

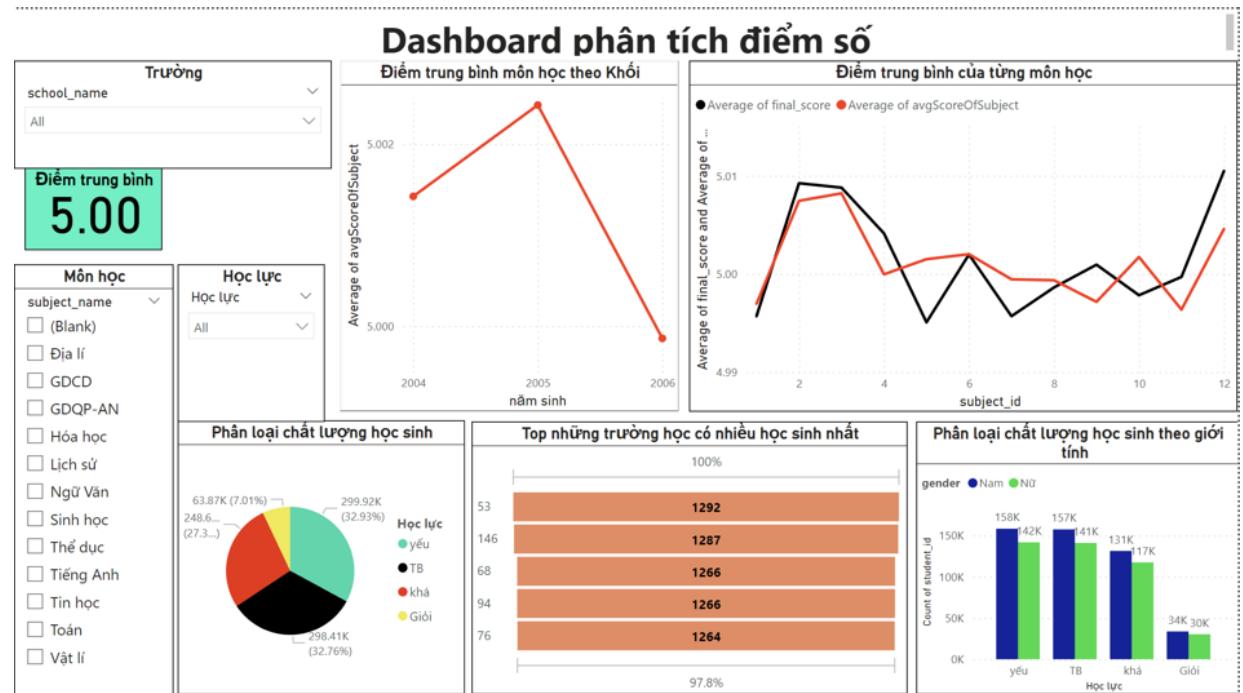
	class_id	fee	level_fee
1	324	780000	Binh Thuong
2	325	760000	Binh Thuong
3	326	830000	Binh Thuong
4	327	620000	Thap
5	328	970000	Cao
6	329	620000	Thap
7	330	750000	Binh Thuong
8	331	940000	Cao
9	332	730000	Thap
10	333	1000...	Cao
11	334	700000	Thap
12	335	740000	Thap
13	336	860000	Cao
14	337	700000	Thap
15	338	640000	Thap
16	339	1000...	Cao
17	340	850000	Cao
18	341	880000	Cao
19	342	750000	Binh Thuong



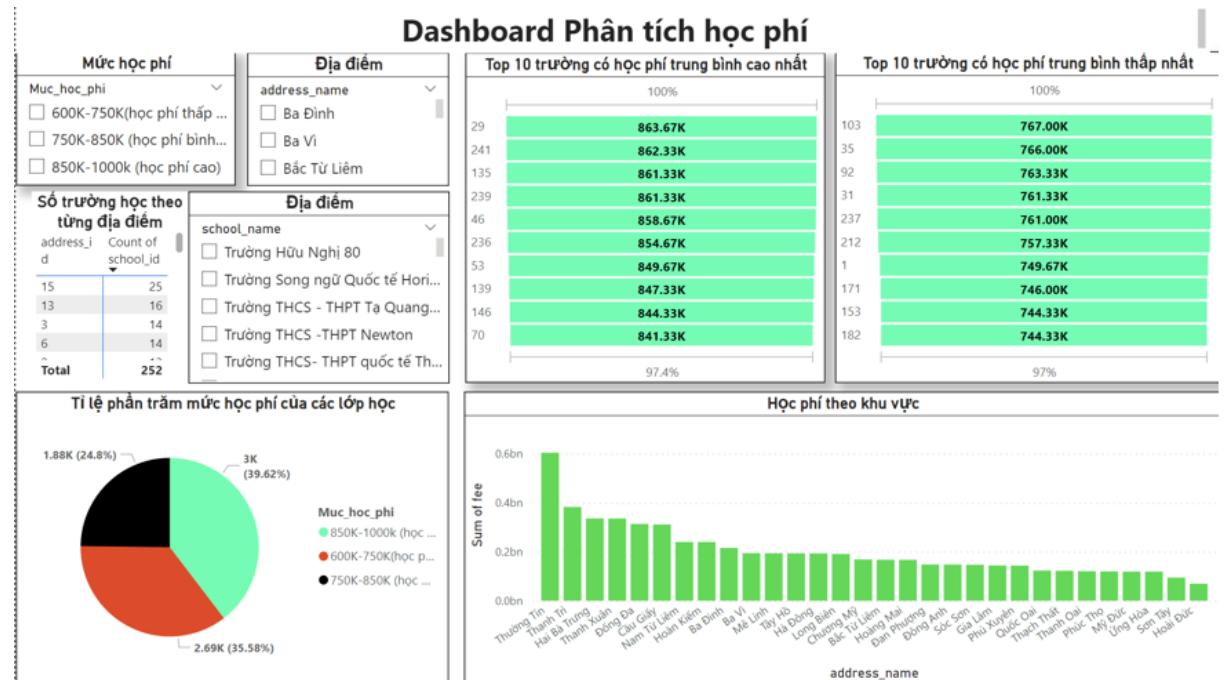
### III. Dashboard (số lượng: 5)

#### 1. Dashboard tổng quan

## 2. Dashboard phân tích điểm số



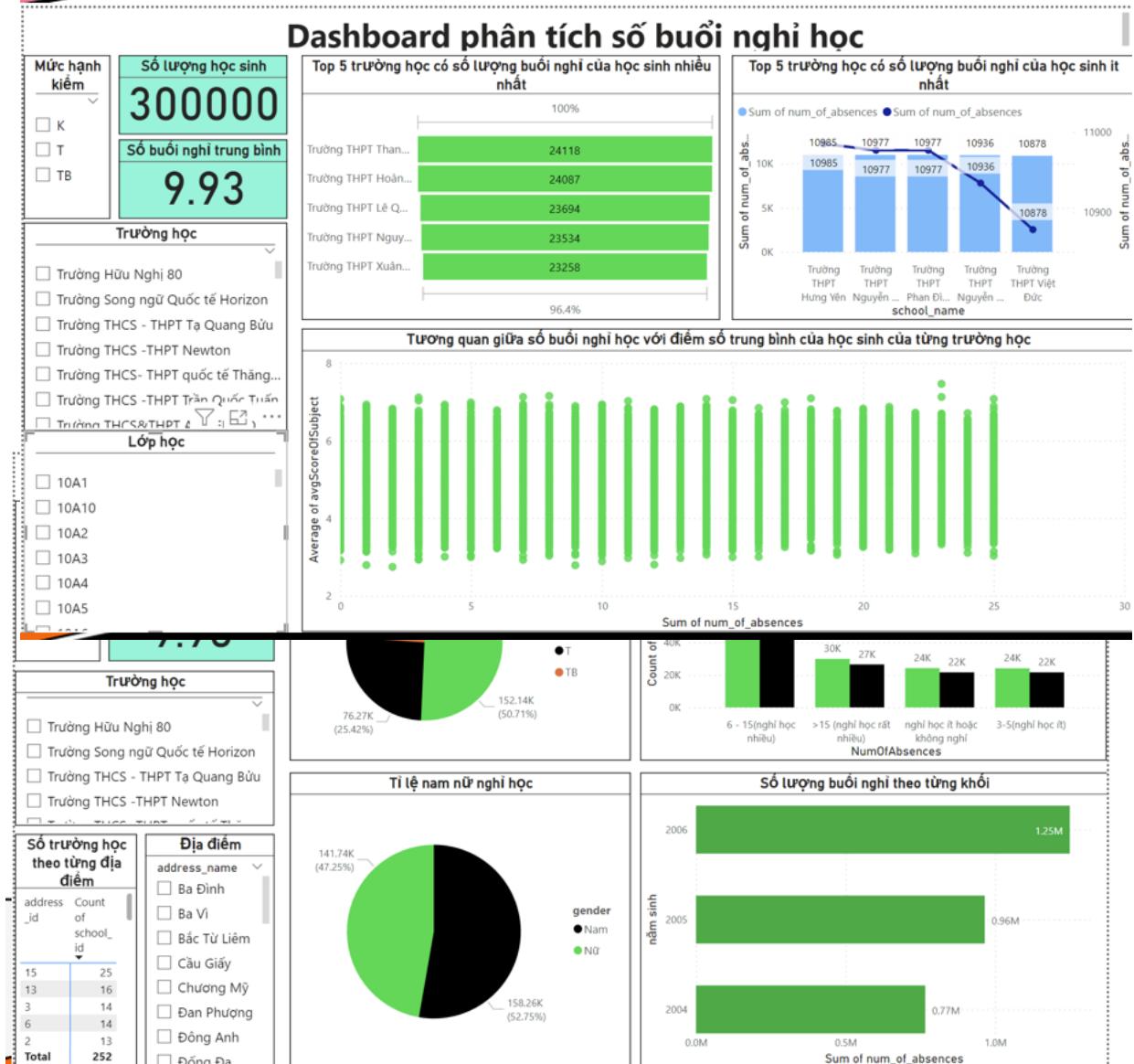
## 3. Dashboard phân tích học phí



#### 4. Dashboard phân tích số buổi nghỉ học

#### 5. Dashboard phân tích tương quan

#### IV. Tổng kết



#### Các nội dung đã thực hiện và đã học được :

- Các kiến thức về mô hình, quy trình hoạt động của hệ thống quản lý giáo dục
- Khảo sát quy trình nghiệp vụ
- Quy trình quản lý dữ liệu, khai thác dữ liệu và phân tích dữ liệu
- Khám phá dữ liệu
- Thiết kế kiến trúc data warehouse
- ETL dữ liệu

- Thiết kế data model OLTP, OLAP
- Xây dựng dashboard phân tích dữ liệu.