ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT
KHOA HỆ THỐNG THÔNG TIN

---ꕥ🕮ꕥ---

BÁO CÁO VÒNG 1

CUỘC THI BUSINESS INTELLIGENCE SEASON 6

Nhóm: VNU

Tp. Hồ Chí Minh, 11/06/2022

THÔNG TIN THÀNH VIÊN

STT	Họ và tên	MSSV	Trường	Ghi chú
1	Nguyễn Thùy Trang	1915588	Trường Đại học Bách Khoa	Nhóm trưởng
2	Phan Phạm Quỳnh Hoa	19521520	Trường Đại Học Công Nghệ Thông Tin	
3	Phạm Thành Đạt	K194111601	Trường Đại học Kinh Tế - Luật	

MỤC LỤC

THÔNG TIN THÀNH VIÊN **1**

MỤC LỤC **2**

PHẦN II: TƯ DUY VỀ KINH DOANH VÀ PHÂN TÍCH DỮ LIỆU **3**

Câu 1. ..... **4**

1.1: Diagram mô tả mối quan hệ giữa 3 bảng dữ liệu. 4

1.2: Thông tin của các khách hàng thuộc phân khúc "Whale" và có mua "Books" trong tháng 5, năm 2021. 5

1.3: Tổng giá trị đơn hàng theo từng ngành hàng và từng phân khúc khách hàng kể từ tháng 8 năm 2021 đến nay 5

1.4: Thông tin của các khách hàng có số lượng đơn hàng cao thứ hai tính trong vị trí địa lý (location) của họ trong tháng 7 năm 2021. 5

1.5: Giá trị đơn hàng đầu tiên của từng khách hàng trong tháng 5 năm 2021 6

1.6: Tính số lượng đơn hàng lũy tiến qua từng ngày trong tháng của từng phân khúc khách hàng. 6

1.7: Thời gian trung bình giữa đơn hàng đầu tiên và đơn hàng thứ hai của các khách hàng trong từng phân khúc khách hàng. 7

Câu 2. ..... **7**

2.1 Gợi ý để xây dựng Data Marts (DM) cho toàn công ty theo nhiều cách, chỉ ra những cách đó và ưu điểm của mỗi cách. 7

2.2 Chọn một kiến trúc phục vụ cho mục đích phân tích liên chi nhánh (cross-branches) thì nên chọn kiến trúc nào? 10

Câu 3. ..... **11**

3.1 Xây dựng quy trình như thế nào? Minh họa lại quy trình đó bằng một sơ đồ 11

3.2 Sử dụng những công nghệ gì để đồng bộ dữ liệu, lý do chọn 12

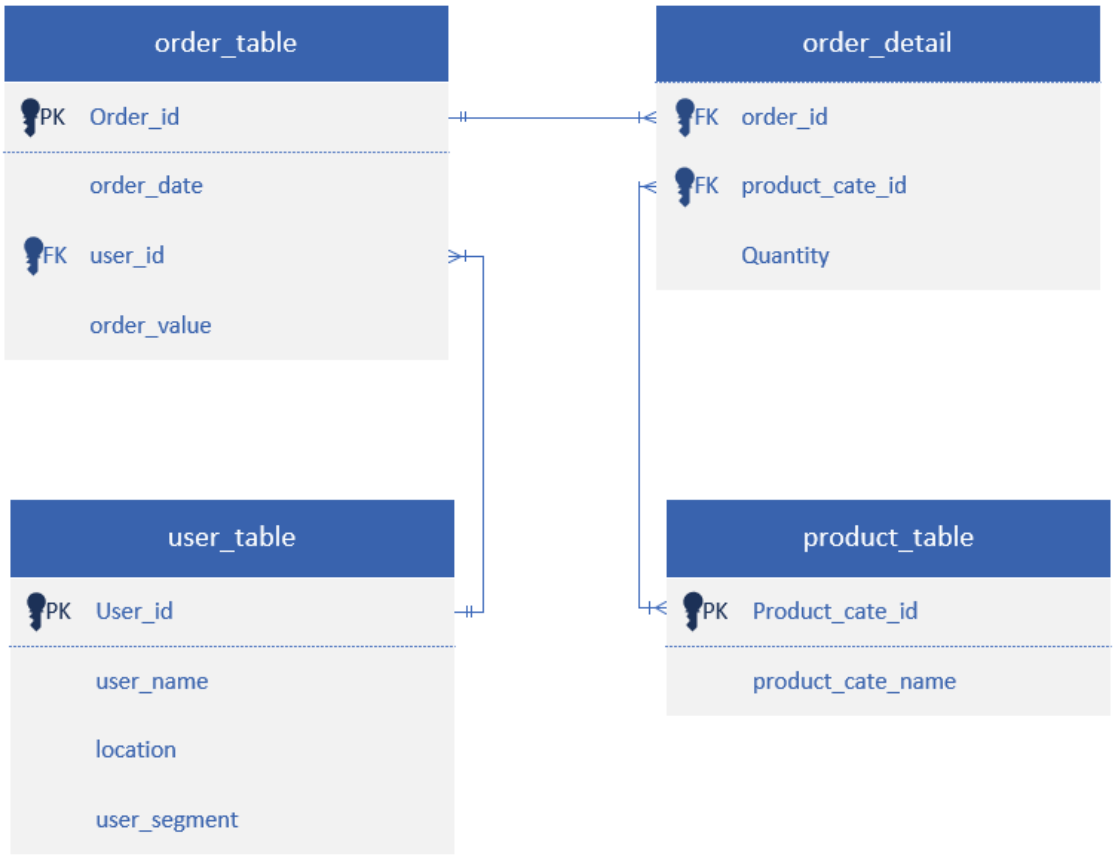
DANH MỤC HÌNH ẢNH, BẢNG BIỂU

Hình 1.1 Diagram mô tả mối quan hệ 3 bảng dữ liệu.....4
Hình 2.1 Mô hình bottom – up design.....8
Hình 2.2 Mô hình top – down design.....9
Hình 2.3 Mô hình hybrid design.....10
Hình 3.1 Diagram minh họa.....12
Bảng 2.1 Bảng so sánh ưu, nhược điểm.....9

PHẦN II: TƯ DUY VỀ KINH DOANH VÀ PHÂN TÍCH DỮ LIỆU

Câu 1.

1.1: Diagram để mô tả mối quan hệ giữa 3 bảng dữ liệu.



Hình 1.1 Diagram mô tả mối quan hệ 3 bảng dữ liệu

```
Create table user_table (
user_id int PRIMARY KEY,
username varchar(100),
location varchar(500),
user_segment varchar(200))

Create table product_table (
product_cate_id int PRIMARY KEY,
product_cate_name varchar(100))

Create table order_table (
order_id int PRIMARY KEY,
order_date date,
user_id int,
order_value float)

Create table order_detail (
order_id int,
```

```
product_cate_id int,
quantity int)
```

1.2: Viết truy vấn SQL để lấy ra thông tin của các khách hàng thuộc phân khúc "Whale" và có mua "Books" trong tháng 5, năm 2021.

```
SELECT user_table.*
FROM user_table,
    order_table,
    order_detail,
    product_table
WHERE user_table.user_id =order_table.user_id
  and product_table.product_cate_id=order_detail.product_cate_id
  and order_detail.order_id= order_table.order_id
  and user_segment = 'Whale'
  and product_cate_name='Books'
  and Month(order_date) = 5
  and YEAR(order_date)=2021;
```

1.3: Viết câu truy vấn SQL để lấy ra tổng giá trị đơn hàng theo từng ngành hàng và từng phân khúc khách hàng kể từ tháng 8 năm 2021 đến nay

```
SELECT SUM(order_value) as "Tong gia tri",
      product_cate_name,
      user_segment
FROM order_table,
    order_detail,
    user_table,
    product_table
WHERE user_table.user_id =order_table.user_id
  and order_table.order_id=order_detail.order_id
  and product_table.product_cate_id = order_detail.product_cate_id
  and MONTH(order_date) >= 8
  and YEAR(order_date) >=2021
GROUP BY product_cate_name,
       user_segment
```

1.4: Viết truy vấn SQL để lấy ra thông tin của các khách hàng có số lượng đơn hàng cao thứ hai tính trong vị trí địa lý (location) của họ trong tháng 7 năm 2021.

```
SELECT TOP 2 WITH TIES COUNT(order_id) AS "SL DON HANG",
              user_table.user_id,
              user_table.location
FROM user_table
join order_table on user_table.user_id = order_table.user_id
WHERE MONTH(order_date) =7
```

```
  and YEAR(order_date)=2021
GROUP BY user_table.location,
        user_table.user_id
ORDER BY [SL DON HANG] DESC
```

1.5: Viết truy vấn SQL để lấy ra giá trị đơn hàng đầu tiên của từng khách hàng trong tháng 5 năm 2021

```
CREATE VIEW min_order AS
  (SELECT order_table.user_id,
         MIN(order_id) AS min_order
   FROM user_table
   join order_table on user_table.user_id = order_table.user_id
   WHERE MONTH(order_date) =5
     and YEAR(order_date)=2021
   GROUP BY order_table.user_id)


SELECT order_value,
       order_table.user_id
FROM order_table,
     min_order
WHERE min_order.min_order = order_table.order_id
```

1.6: Viết truy vấn SQL để tính số lượng đơn hàng lũy tiến qua từng ngày trong tháng của từng phân khúc khách hàng.

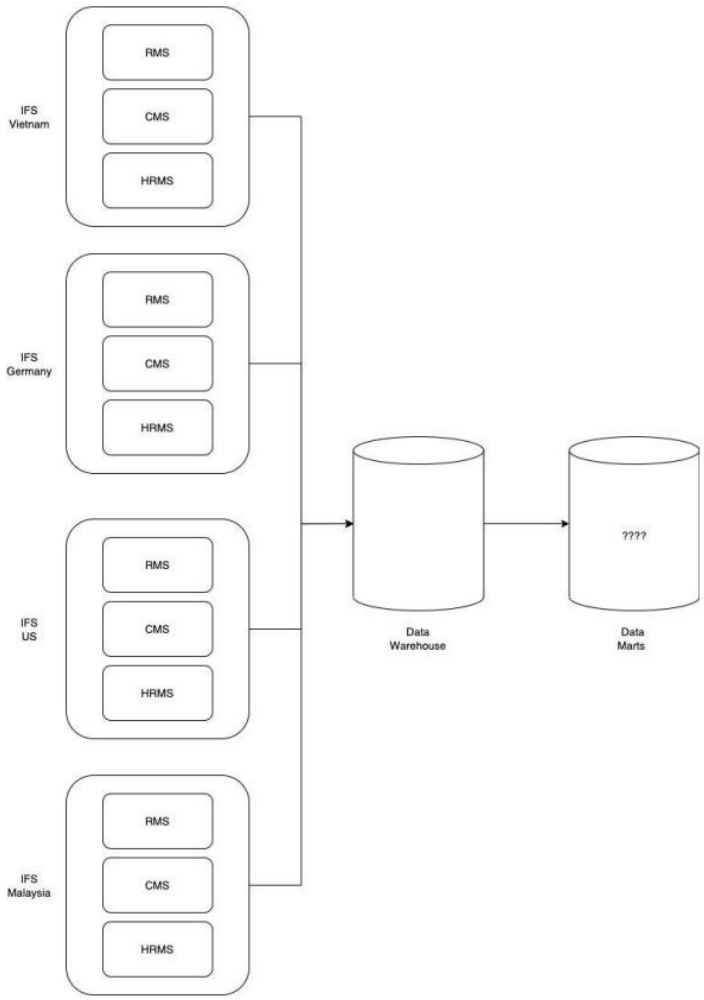
```
CREATE VIEW cau6 AS
SELECT COUNT(order_id) AS "SL",
       DAY(order_date) AS DAY,
       MONTH(order_date) AS MONTH,
       user_segment
FROM order_table,
     user_table
WHERE user_table.user_id =order_table.user_id
GROUP BY DAY(order_date),
         user_segment,
         MONTH(order_date)


SELECT *,
       SUM(SL) OVER (PARTITION BY MONTH
                     ORDER BY DAY) AS RunningTotal
FROM cau6
```

1.7: Viết truy vấn để tính được thời gian trung bình giữa đơn hàng đầu tiên và đơn hàng thứ hai của các khách hàng trong từng phân khúc khách hàng.

Câu 2.

Mỗi chi nhánh đều có một team Business Intelligence (BI) hoạt động độc lập, và ở trụ sở chính tại Việt Nam cũng có một team BI phụ trách cho trụ sở chính và hỗ trợ cho các chi nhánh khác.

Hiện tại bạn đang làm Business Intelligence Manager tại trụ sở chính ở Việt Nam, ban giám đốc đang cần bạn đưa ra những gợi ý để xây dựng Data Marts cho toàn công ty theo nhiều cách khác nhau, hãy chỉ ra những cách đó và ưu điểm của mỗi cách. Giả sử ban giám đốc muốn chọn một kiến trúc phục vụ cho mục đích phân tích liên chi nhánh (cross-branches) thì nên chọn kiến trúc nào?

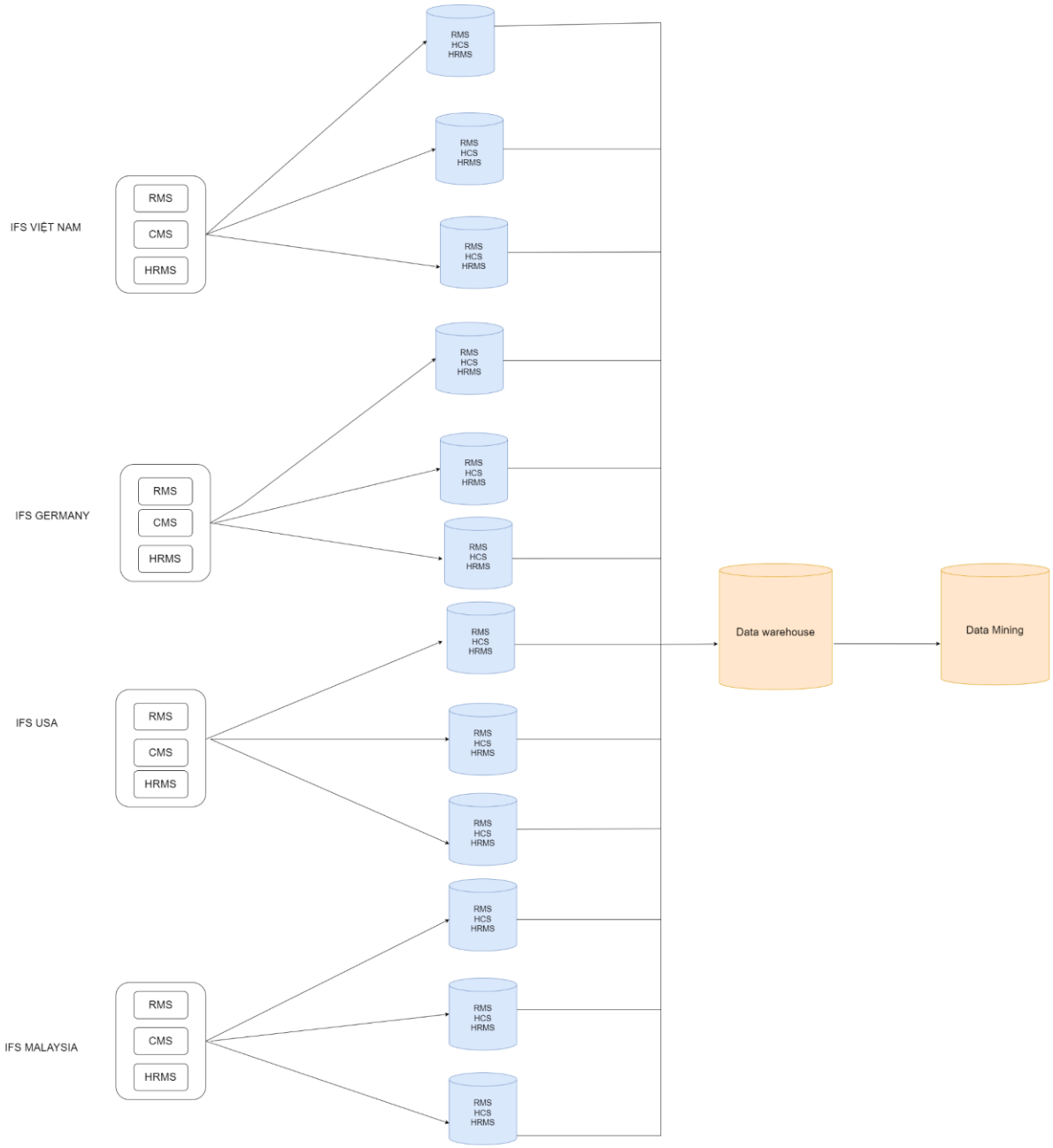


2.1 Ban giám đốc đang cần bạn đưa ra những gợi ý để xây dựng Data Marts (DM) cho toàn công ty theo nhiều cách khác nhau, hãy chỉ ra những cách đó và ưu điểm của mỗi cách.

a. Bottom – up design:

Trong cách tiếp cận này, mỗi DM sẽ là một hệ thống riêng lẻ (RMS, CMS, HRMS), được tạo để cung cấp khả năng báo cáo và phân tích cho các quy trình kinh doanh của công ty. Sau đó, các DM này sẽ được tích hợp để tạo ra một kho dữ liệu – data warehouse (DW) toàn diện.

Việc tích hợp các DM này được thực hiện theo cấu trúc mạch nối (bus architecture) của DW. Trong kiến trúc này, một thứ nguyên (dimension) được chia sẻ giữa các dữ kiện trong hai hoặc nhiều DM và được tích hợp từ các DM để xây dựng được DW.

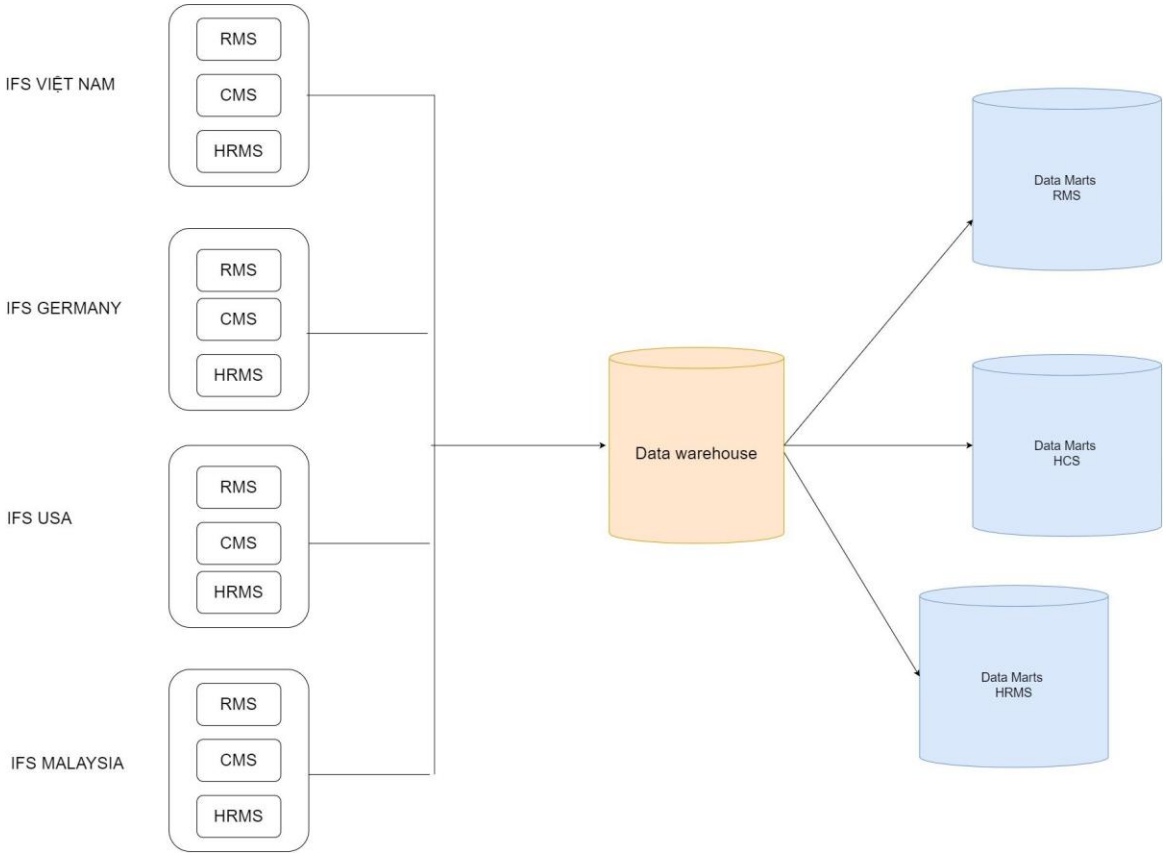


Hình 2.1 Mô hình bottom – up design

b. *Top – down design:*

Trong cách tiếp cận này, DW được xây dựng đầu tiên. Sau đó, các DM được tạo ra từ DW.

Cách tiếp cận từ trên xuống được thiết kế bằng cách sử dụng mô hình dữ liệu doanh nghiệp đã được chuẩn hóa. Dữ liệu "atomic", tức là dữ liệu ở mức độ chi tiết thấp nhất, được lưu trữ trong DW. Các DM chứa dữ liệu cần thiết cho các hệ thống RMS, CMS, HRMS được tạo từ DW.



Hình 2.2 Mô hình top – down design

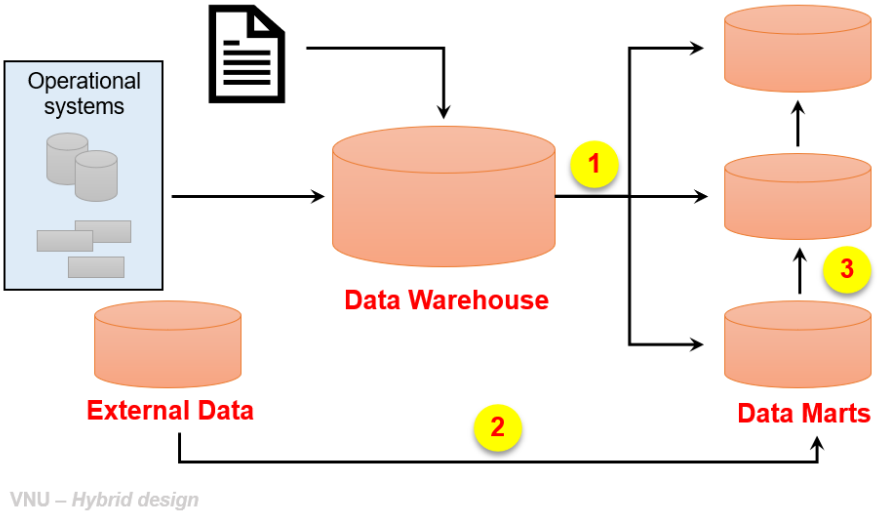
	Ưu điểm	**Nhược điểm**
Bottom – up design	+ Chứa các DM nhất quán và các DM này có thể được phân chuyển rất nhanh. + Vì các DM được xây dựng trước nên có thể xuất những báo cáo một cách nhanh chóng. + Có thể mở rộng DW dễ dàng để đáp ứng các đơn vị kinh doanh mới bằng việc tạo ra các DM mới và tích hợp chúng với các DM trước đó.	+ Vị trí của DM và DW bị đảo ngược.
Top – down design	+ Cung cấp chế độ xem dữ liệu theo chiều nhất quán giữa các DM vì tất cả các DM đều được tải từ DW. + Chủ động và có thể linh hoạt trong điều chỉnh khi thị trường có sự thay đổi. Vì có thể tạo những DM mới từ DW rất dễ dàng.	+ Không linh hoạt khi có sự thay đổi nhu cầu của các bộ phận trong giai đoạn thực hiện. + Đại diện cho một dự án rất lớn và chi phí thực hiện dự án là đáng kể.

Bảng 2.1 Bảng so sánh ưu, nhược điểm

c. Hybrid design:

Trong cách tiếp cận này, cho phép các DM kết hợp dữ liệu từ ngoài DW. Điều này phù hợp cho nhiều trường hợp, đặc biệt là trong trường hợp muốn tích hợp đặc biệt (Adhoc), chẳng hạn như thêm mới nhóm hoặc sản phẩm vào tổ chức.

Cách tiếp cận này có thể được xem là tốt nhất, phù hợp với nhiều môi trường cơ sở dữ liệu và quay vòng triển khai nhanh cho mọi tổ chức. Nó cũng đòi hỏi nỗ lực làm sạch dữ liệu ít nhất. Hybrid DM cũng hỗ trợ các cấu trúc lưu trữ lớn và phù hợp nhất để có thể linh hoạt cho các ứng dụng cần ít sự tập trung vào dữ liệu hơn.



Hình 2.3 Mô hình hybrid design

2.2 Giả sử ban giám đốc muốn chọn một kiến trúc phục vụ cho mục đích phân tích liên chi nhánh (cross-branches) thì nên chọn kiến trúc nào?

a. So sánh:

- *Top – down:* sẽ cần phân tích nhu cầu kinh doanh tổng thể các chi nhánh, lập kế hoạch cách phát triển, thiết kế và triển khai DW. Quy trình này đầy hứa hẹn, nó sẽ đạt được kết quả xuất sắc vì **dựa trên bức tranh tổng thể** về mục tiêu cần đạt được. Và về nguyên tắc, nó đảm bảo DW được tích hợp tốt và nhất quán như ưu điểm được nêu ở trên. Tuy nhiên, có một vài những nhược điểm nếu xây dựng theo kiến trúc này:
  - Dự toán chi phí cao với thời gian thực hiện dài hạn → không khuyến khích các nhà quản lý công ty bắt tay vào những dự án kiểu này.
  - Phân tích và tập hợp tất cả các nguồn ở các hệ thống con (RMS, CMS, HRMS) là một việc rất khó khăn, cũng bởi vì không có khả năng cao là tất cả chúng đều có sẵn và ổn định cùng một lúc (công ty có nhiều chi nhánh).
  - Dự báo nhu cầu cụ thể của các bộ phận tham gia là vô cùng khó khăn → có thể khiến quá trình phân tích đi vào bế tắc.
  - Vì không có nguyên mẫu (prototype) nào sẽ được chuyển giao trong thời gian ngắn hạn, công ty không thể kiểm tra xem dự án phân tích này có hữu ích, hợp lý hay không.

- *Bottom – up*: DW được xây dựng từng bước và một số kho dữ liệu được tạo lặp đi lặp lại. Nếu cách tiếp cận này kết hợp với việc tạo mẫu nhanh, thời gian và chi phí

cần thiết để thực hiện có thể giảm đáng kể đến mức các nhà quản lý công ty sẽ nhận thấy dự án phân tích của họ đang được thực hiện hữu ích như thế nào.
Cách tiếp cận này không có rủi ro, vì nó **có được bức tranh tổng thể** về toàn bộ lĩnh vực ứng dụng.

b. Kết luận:

Nhóm đề xuất xây dựng theo kiến trúc Bottom – up.

Vì từng chi nhánh đã có các hệ thống DM riêng và bên cạnh các ưu điểm như, các DM này có thể được phân chuyển rất nhanh, có thể xuất những báo cáo một cách nhanh chóng, có thể mở rộng DW dễ dàng để đáp ứng các đơn vị kinh doanh mới bằng việc tạo ra các DM mới và tích hợp chúng với các DM trước đó, thì kiến trúc này cũng tốt cho việc phân tích cục bộ, tiết kiệm chi phí cũng như triển khai nhanh.

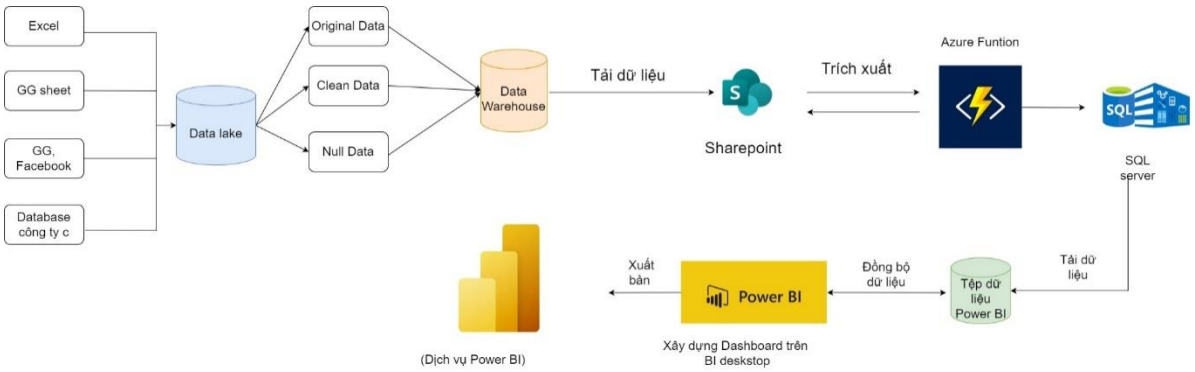
Từ đó, các DM của các chi nhánh sẽ tập trung lại thành 1 DW lớn, từ đây sẽ triển khai tiếp tục việc phân tích liên chi nhánh.

→ *Thực hiện kiến trúc Bottom – up giúp tiết kiệm thời gian, chi phí cũng như scope của dự án không quá lớn sẽ dễ thành công hơn.*

Câu 3.

A là một công ty chuyên cung cấp các giải pháp về Business Intelligence. Gần đây, A có nhận được một dự án của một khách hàng là công ty B làm trong lĩnh vực bán lẻ, chuyên cung cấp các mặt hàng về chăm sóc da và làm đẹp. Trước đây, khi mới thành lập, số lượng đơn hàng không nhiều nên công ty B chủ yếu lưu lại thông tin giao dịch và thông tin khách hàng riêng lẻ trên nhiều file excel và google sheet. Các dữ liệu thống kê về quảng cáo thì team Marketing của B đang hằng ngày tải về từ Google và Facebook dưới dạng file *.csv. Sau một năm, công ty B chuyển sang sử dụng phần mềm quản lý bán hàng của công ty C và các thông tin về quảng cáo, giao dịch, khách hàng được lưu trên hệ thống mà công ty C cung cấp. Hằng ngày, đội ngũ kỹ thuật của công ty C sẽ lấy dữ liệu về từ hệ thống của C về và lưu vào database riêng của công ty (đang chạy trên nền tảng Microsoft SQL Server). Để tiện cho việc lưu trữ tài nguyên dữ liệu và phân tích phục vụ cho việc phát triển, công ty B yêu cầu bên công ty của A đồng bộ dữ liệu từ các nguồn khác nhau về một nguồn từ đó xây dựng các báo cáo phân tích dựa trên nguồn dữ liệu thống nhất đó. Nếu bạn là team leader của dự án này tại công ty A:

3.1 Team của bạn sẽ xây dựng quy trình như thế nào? Minh họa lại quy trình đó bằng một sơ đồ (diagram).



Hình 3.1 Diagram minh họa

Quy trình được thực hiện như sau:

- Thu thập dữ liệu từ nhiều nguồn khác nhau và thực hiện quá trình **SSIS** như sau:
  Bước 1 - Clean Data Lake and Load Dataset: Tiến hành đưa toàn bộ dữ liệu thô vào bảng tương ứng trong Data Lake, thực hiện tiền xử lý dữ liệu và đổ vào các bảng dữ liệu tương ứng
  Bước 2 – ETL data to Dimension Table: Thực hiện đổ dữ liệu vào các bảng Dimension trong Data warehouse
  Bước 3 – Fact: Thực hiện đổ dữ liệu vào bảng Fact trong Data warehouse
  Bước 4 – Set Foreign Keys: Cài đặt các khóa ngoại tương ứng theo thiết kế lược đồ
- Sau khi xây dựng kho dữ liệu thành công, dữ liệu nghiệp vụ được tải lên hệ thống SharePoint (một nền tảng chia sẻ thông tin của Microsoft) và phía người dùng sẽ cung cấp API để đội ngũ triển khai BI có thể trích xuất dữ liệu từ hệ thống.
- Tiếp đến quy trình ETL để trích xuất dữ liệu, biến đổi và nạp vào cơ sở dữ liệu tạm (SQL Server) với mã Python và nền tảng Azure Functions, sau đó tải vào tập dữ liệu Power BI.
- Đồng bộ dữ liệu và xây dựng hệ thống Dashboard trên Power BI Desktop để người dùng sử dụng.
- Quá trình đồng bộ dữ liệu xong sẽ được xuất bản cho công ty B sử dụng.

3.2 Team bạn sẽ sử dụng những công nghệ gì để đồng bộ dữ liệu, lý do chọn những công nghệ đó

- Công nghệ đồng bộ hóa mà nhóm đưa ra cho tình huống của doanh nghiệp trên đó là sử dụng công nghệ **Cloud BI** thông qua phương pháp khảo sát hiện trạng dữ liệu, quản lý và ra quyết định tại doanh nghiệp sẽ mang lại hiệu quả cho việc lưu trữ tài nguyên dữ liệu và phân tích phục vụ cho việc phát triển. Sau đó, nghiên cứu áp dụng phương pháp thực nghiệm để triển khai giải pháp BI trên nền tảng Microsoft Azure, nhờ giải pháp tiếp cận nguồn dữ liệu từ cổng API giúp cho giải pháp trở nên linh hoạt, dễ dàng chuyển đổi nhà cung cấp.
- Dùng ETL Pentaho một công cụ miễn phí để ETL data từ nhiều nguồn.
- Dùng Microsoft SQL server để đồng bộ hóa và sử dụng Microsoft Power Bi để làm báo cáo.

Cloud BI là gì? Tại sao nên chọn giải pháp này?

- ***Cloud BI*** là nền tảng phân tích dữ liệu dựa trên Looker- công cụ thu thập, chuyển đổi dữ liệu hiện có của Doanh nghiệp thành các tri thức kinh doanh (BI) như tạo ra các báo cáo phân tích từ vô số mô hình dữ liệu khác nhau cho đến tích hợp quy trình làm việc với các ứng dụng dữ liệu tùy chỉnh.
- ***Microsoft Azure*** là một nền tảng điện toán đám mây, cho phép người dùng truy cập và quản lý các dịch vụ cũng như tài nguyên do Microsoft cung cấp
- ***Kiểm soát và an toàn dữ liệu:*** Sử dụng Cloud BI với cấu trúc dữ liệu có thể vận dụng sức mạnh của công nghệ data warehouse và truy vấn cơ sở dữ liệu trong thời gian thực để dữ liệu luôn chính xác và mới, giúp việc đồng bộ hóa dữ liệu trở nên nhanh chóng và hiệu quả. Sau khi dữ liệu đồng bộ hóa và lưu trữ tập trung trên Cloud sẽ luôn được cập nhật phiên bản mới nhất khi có thay đổi và ghi nhận thời gian thực.
- ***Tiết kiệm và linh hoạt*****:** Dữ liệu lưu trữ lớn và linh hoạt cho doanh nghiệp khi muốn thu hẹp/mở rộng phạm vi quản lý dữ liệu tùy theo giai đoạn thực tế. Với khả năng khôi phục dữ liệu sau thảm họa, giúp tiết kiệm thời gian, tránh lãng phí đầu tư và tạo uy tín của bên thứ ba. Giảm chi phí cho phần cứng, chỉnh sửa và chia sẻ dữ liệu chỉ cần có internet.
- ***Ngoài Cloud Bi,*** sử dụng kết hợp Microsoft SQL Server, Microsoft Power Bi có sự phối hợp liên kết với nhau cùng với đó là chi phí vô cùng hợp lý.