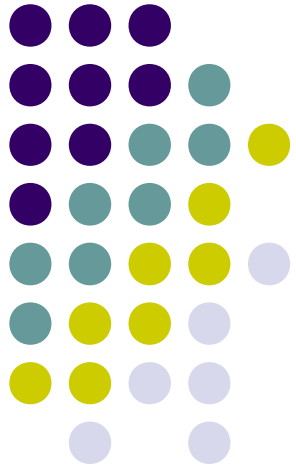
Tìm hiểu về DW 2.0
Chương 19, 20, 21

Thành viên trình bày:

1041117	Hứa Chấn Quốc
1041357	Nguyễn Thành Khang
1041311	Lê Hoàng Minh Châu



Chương 19 : DW 2.0 & unstructured data

Nội dung chính:

1) Khái niệm unstructured data
2) Xử lý văn bản phi cấu trúc
   - Phương pháp thực hiện
   - Tích hợp văn bản
3) Cách sử dụng

1/ Khái niệm unstructured data

- Là 1 dạng dữ liệu trong data warehouse có nguồn gốc từ unstructured text (txt, xls, pdf, csv,…).
- Dùng unstructured text sẽ cho kết quả phân tích sai.
- Để chuyển từ unstructured text thành unstructured data thì qua các bước:

1) Đọc văn bản
2) Tích hợp văn bản

2/ Xử lý văn bản phi cấu trúc –

Phương pháp thực hiện

- Con người tự làm
- Xử lý bằng công cụ có sẵn : textual ETL

=> cho kết quả tốt nhất

2/ Xử lý văn bản phi cấu trúc –
Tích hợp văn bản

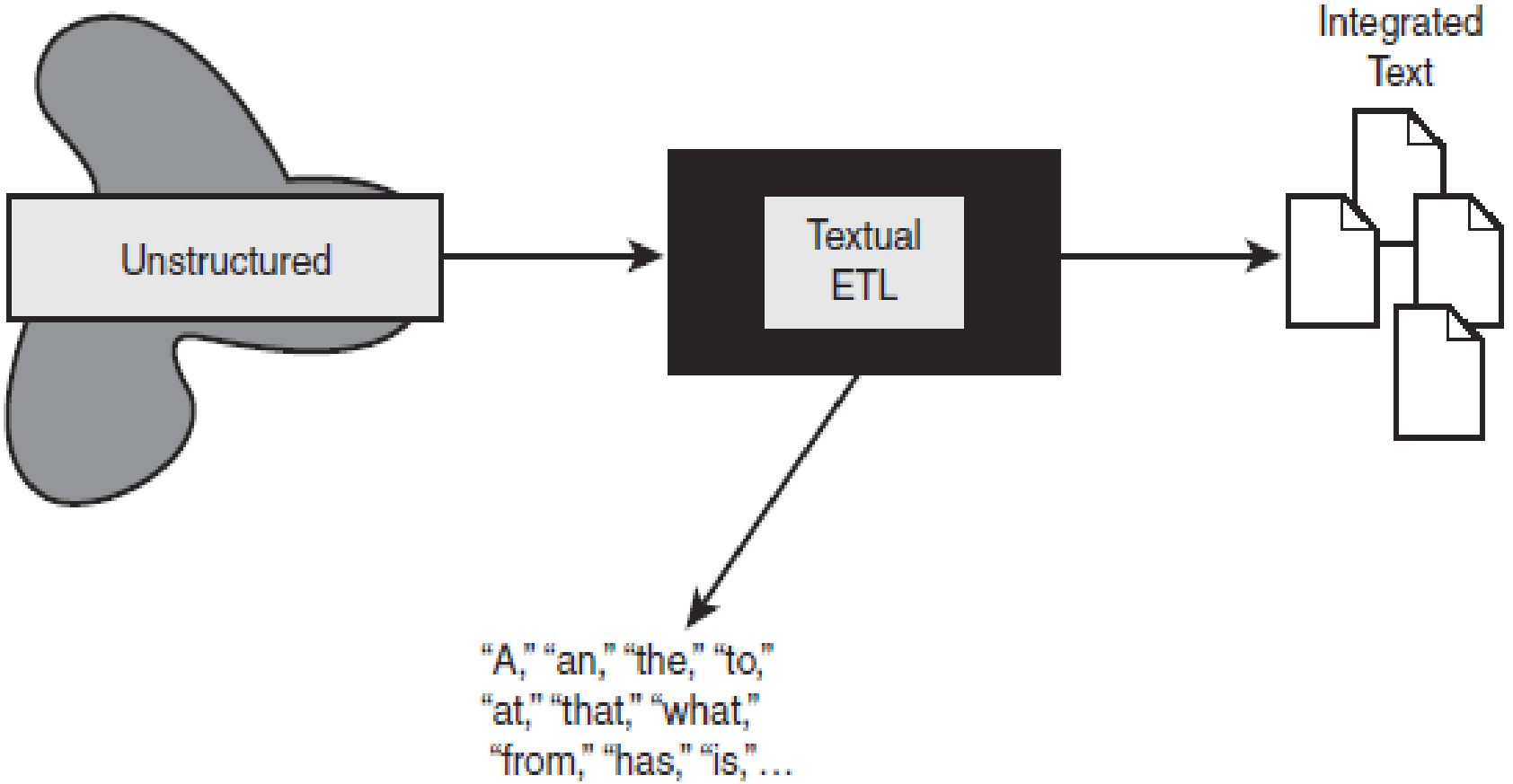
- ✓ Simple editing : chuyển mọi ký tự hoa thành thường và bỏ mọi dấu câu.

Lincoln stood and said - “Four score and seven years ago, our forefathers”

lincoln stood and said four score and seven years ago our forefathers

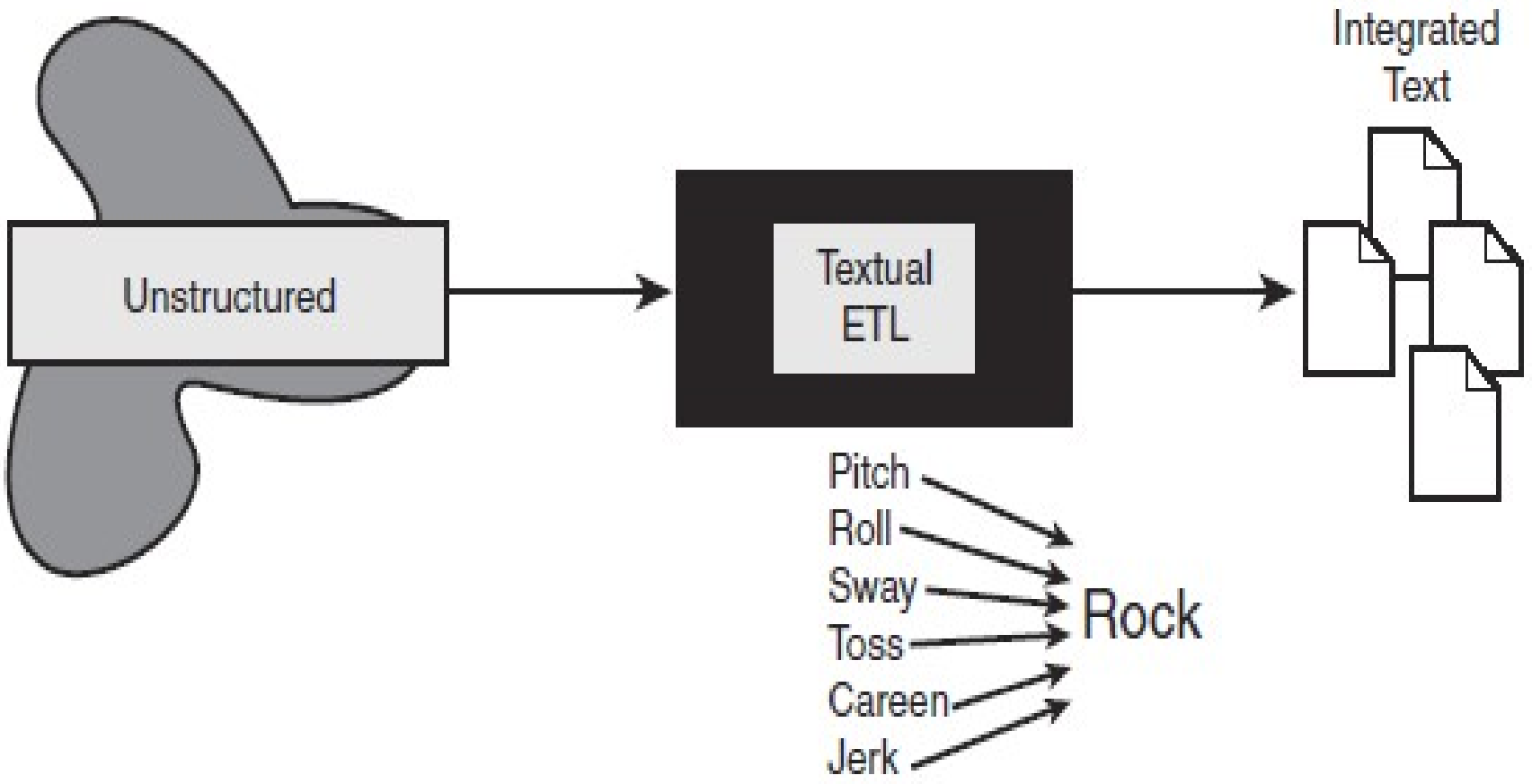
2/ Xử lý văn bản phi cấu trúc –
Tích hợp văn bản

- Stop-word removal : Loại bỏ mọi loại từ ngoại trừ danh từ.



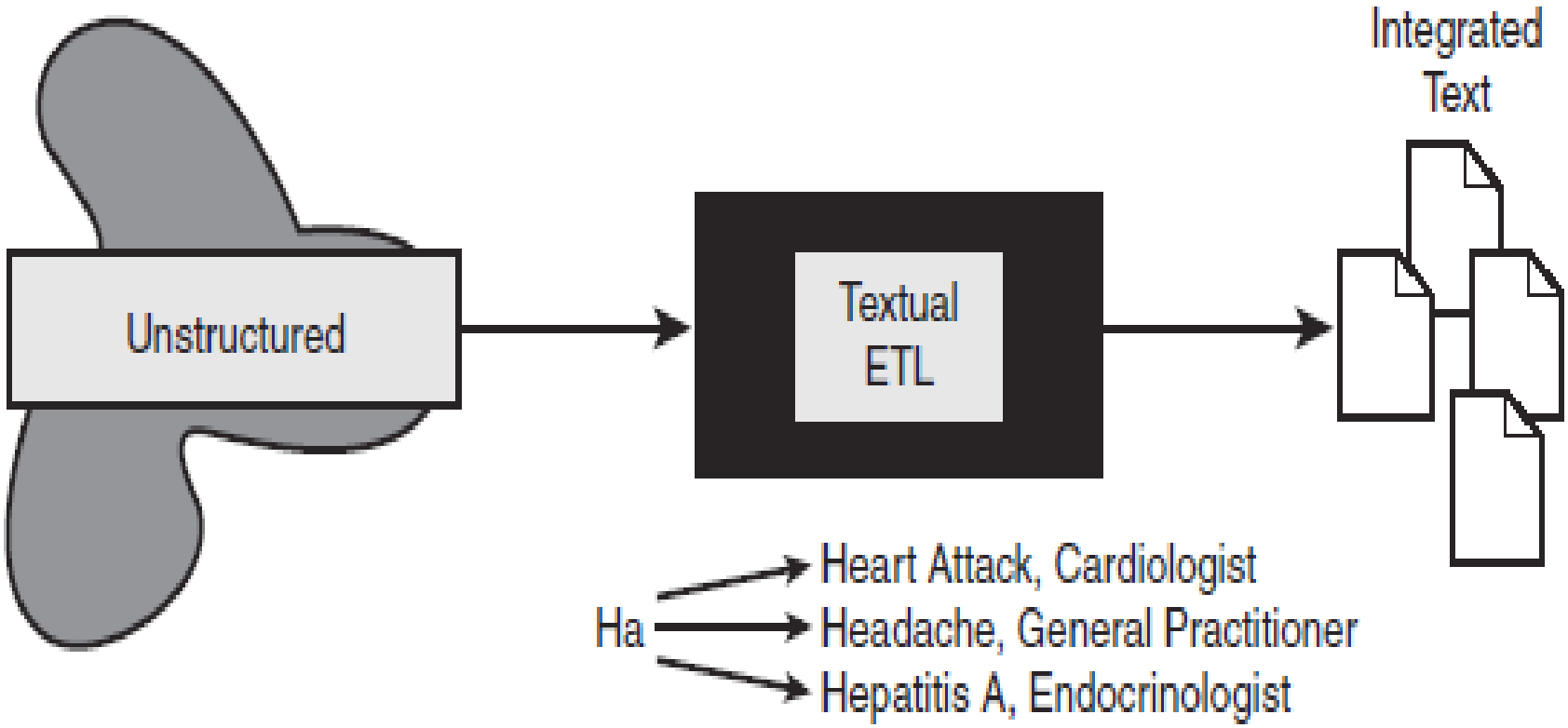
2/ Xử lý văn bản phi cấu trúc –
Tích hợp văn bản

- ✓ Synonym replacement : thống nhất các từ đồng nghĩa bằng 1 từ thông dụng nhất.



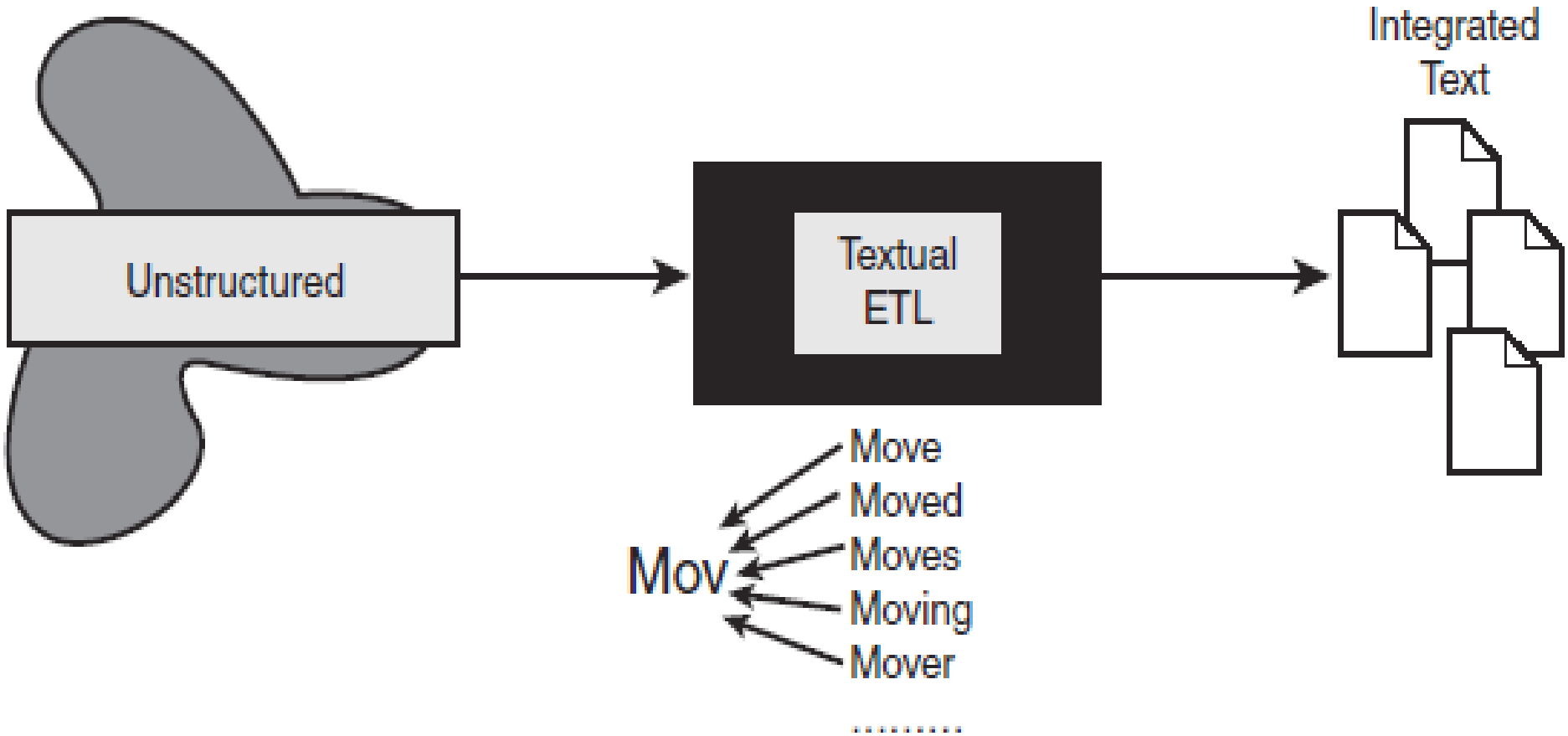
2/ Xử lý văn bản phi cấu trúc –
Tích hợp văn bản

- ✓ Homographic resolution : làm rõ nghĩa những từ có ý nghĩa khác nhau.



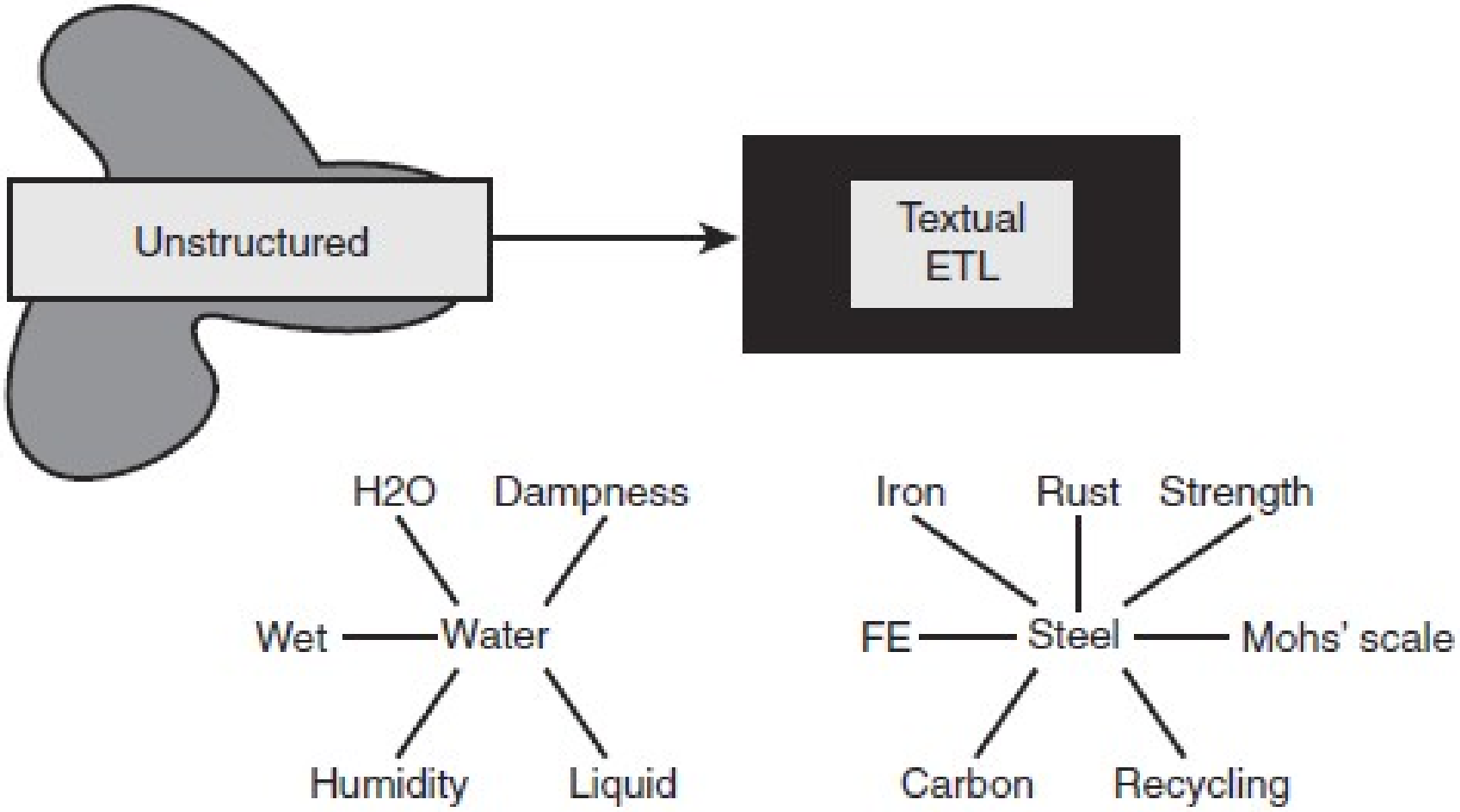
2/ Xử lý văn bản phi cấu trúc –
Tích hợp văn bản

- ✓ Stemming : chuyển các từ về từ gốc latin



2/ Xử lý văn bản phi cấu trúc –
Tích hợp văn bản

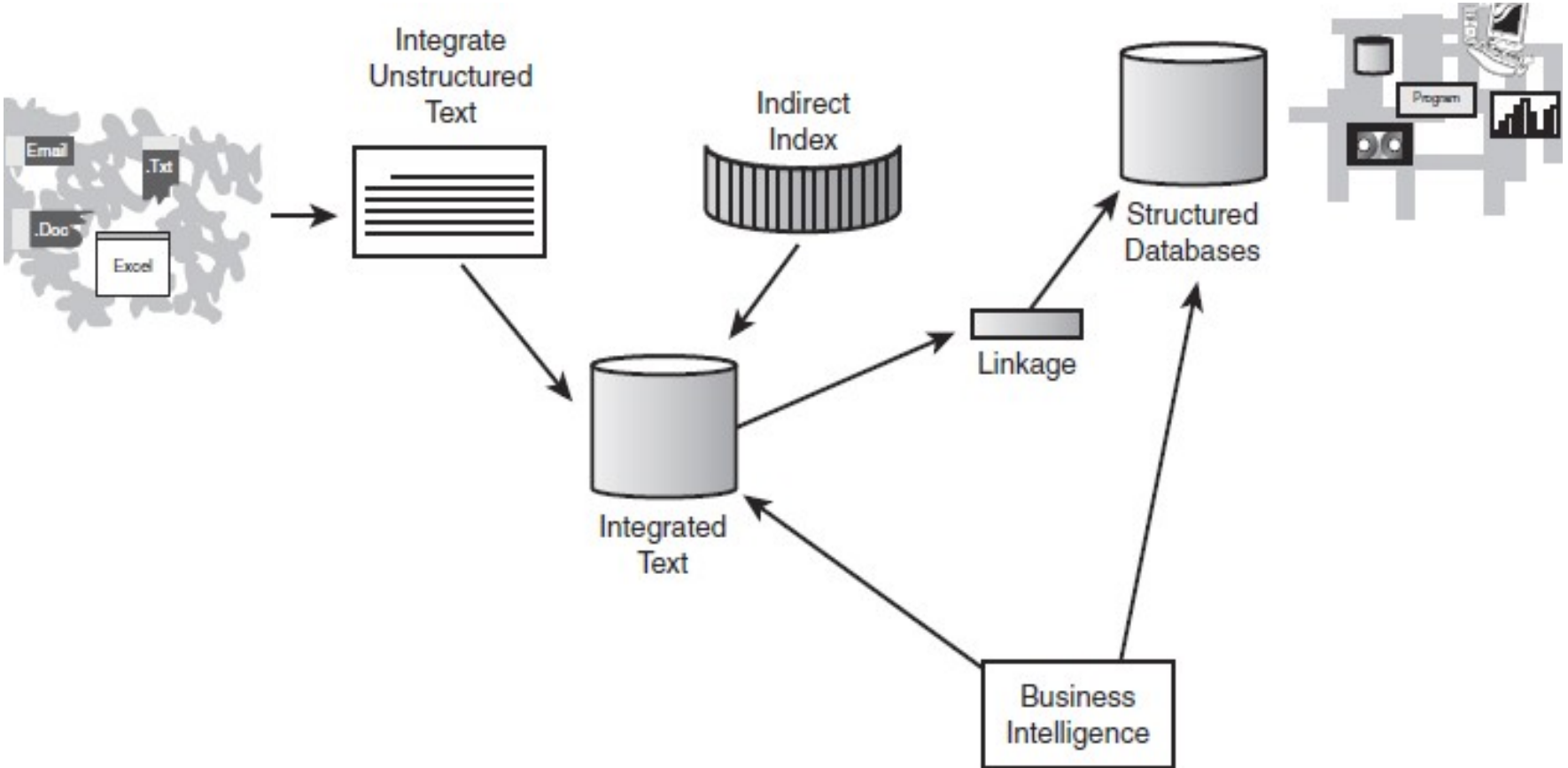
- ✓ Creating themes : Gom nhóm các từ theo 1 chủ



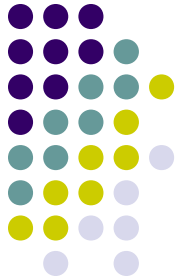
3/ Cách sử dụng

Đưa unstructured data vào relational database để được :

- Phân tích bằng BI
- Tìm kiếm trực tiếp hoặc gián tiếp
- Kết nối với CSDL có cấu trúc để thực hiện các truy vấn phức tạp



Chương 20:
DW 2.0 & The system of record



Nội dung chính:

1. khái niệm
2. Mapping data
3. Nguồn dữ liệu khác

1. Khái niệm

- The system of record là các nguồn dữ liệu tốt nhất của data warehouse .
- Các nguồn dữ liệu có thể dùng cho DW tồn tại trong operational legacy environment dưới dạng chương trình ứng dụng, báo cáo, tập tin, cơ sở dữ liệu .

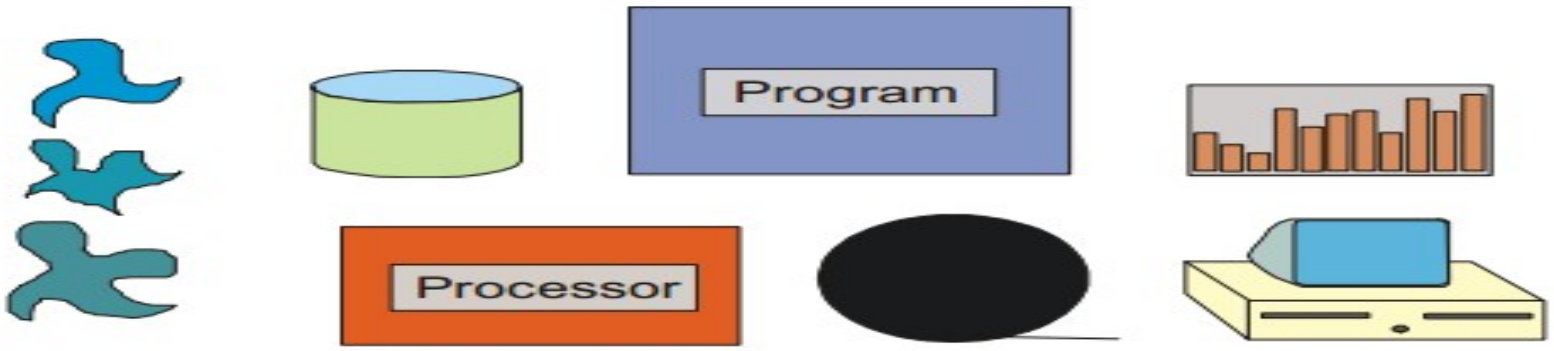


Fig sor.2
The elements of the operational environment

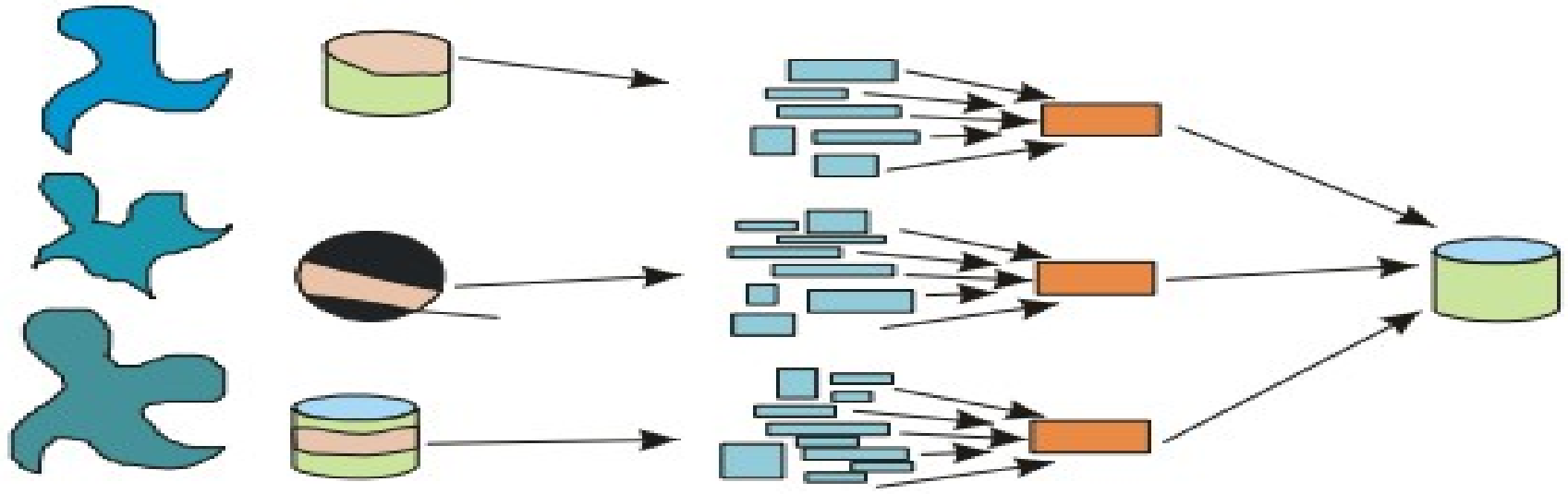
1. Khái niệm

Đặt điểm Của dữ liệu tốt

- Chính xác nhất
- Hoàn thiện nhất
- Mới nhất
- Đáng tin cậy
- Truy cập nhiều nhất

2. Mapping data

Sau khi đã chọn được các nguồn dữ liệu tốt nhất thì phải chuyển hóa chúng về 1 nguồn dữ liệu đích (target data)



2. Mapping data

vài ví dụ về chuyển hóa dữ liệu

Fig sor.13
A simple mapping

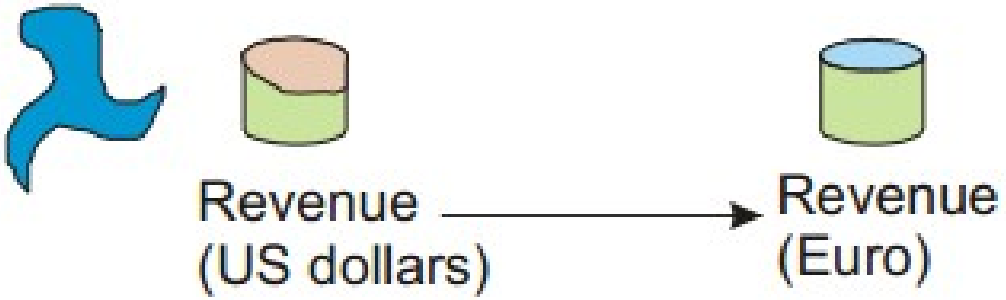
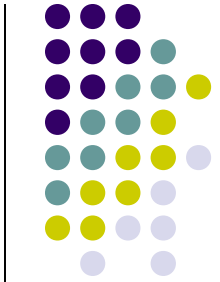


Fig sor.14
A more complex mapping

3. Nguồn dữ liệu khác



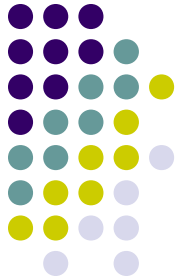
- Data mart có thể rút trích dữ liệu từ mọi khu vực trong DW :interactive, Integrated, Near Line và Archival.
- Tất cả dữ liệu đó đều là nguồn dữ liệu đầu vào của data mart

Chương 21 : Miscellaneous topics

Các khái niệm cơ bản

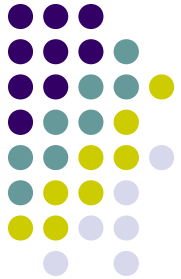
- Kho dữ liệu (Data Warehouse - DW)
- Kho dữ liệu cục bộ (Data Mart - DM)
- Data mart phụ thuộc (Dependent Data Mart)
- Data mart độc lập (Independent Data Mart)

Kho dữ liệu (Data Warehouse - DW)



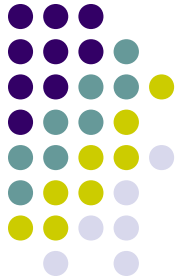
- *Kho dữ liệu là tuyển tập các cơ sở dữ liệu tích hợp, hướng chủ đề, được thiết kế để hỗ trợ cho chức năng trợ giúp quyết định*

. Kho dữ liệu cục bộ (Data Mart - DM)



- Kho dữ liệu cục bộ là CSDL có những đặc điểm giống với kho dữ liệu nhưng với quy mô nhỏ hơn và lưu trữ dữ liệu về một lĩnh vực, một chuyên ngành

Data mart phụ thuộc (Dependent Data Mart):



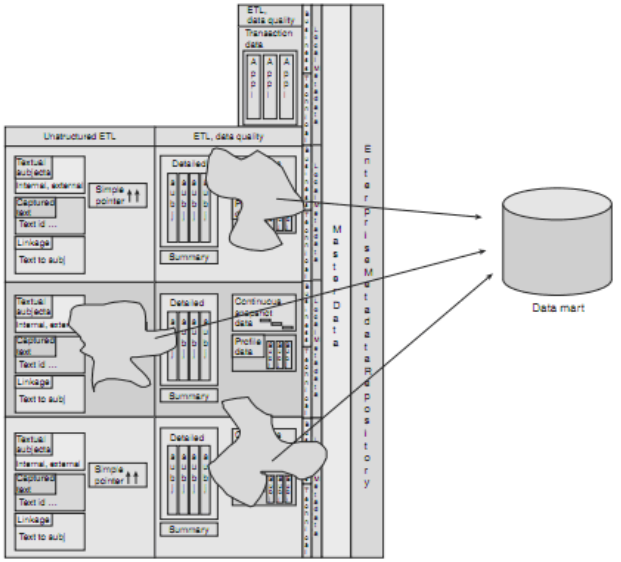
- Chứa những dữ liệu được lấy từ DW và những dữ liệu này sẽ được trích lọc và tinh chế, tích hợp lại ở mức cao hơn để phục vụ một chủ đề nhất định của Datamart

Data mart độc lập (Independent Data Mart)

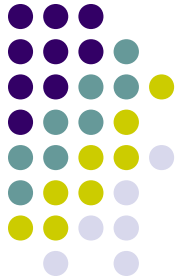
- Không giống như Datamart phụ thuộc, Data mart độc lập được xây dựng trước DW và dữ liệu được trực tiếp lấy từ các nguồn khác nhau

Hình vẽ

- Mô tả về hệ thống mới



Lợi ích của data mart



- Đưa ra những thông tin , cấu trúc mà con người muốn tìm nhanh chóng, chính xác
- Giảm chi phí thực hiện dữ liệu khi lấy thông tin ra khỏi kho dữ liệu
- khi di chuyển dữ liệu đến máy khác, chu kỳ máy 2.0 DW doanh nghiệp môi trường kho dữ liệu được bảo tồn

Chuyển dữ liệu:

- Data mart tổng hợp , lấy dữ liệu từ nhiều nguồn khác nhau do đó việc chuyển đổi dữ liệu từ các định dạng khác nhau từ các nguồn khác nhau về 1 cái gì thống nhất với nhau và nó được lưu trữ trong data mart để phục vụ cho công việc và chia sẻ kho dữ liệu đó tới người dùng cuối.

GIÁM SÁT DW 2.0

- Khi có 1 hành động bên trong data mart tiến hành truy vấn để lây thông tin và muốn xem những thông tin thì sẽ sinh ra các data mart, nên chúng ta cần giám sát trường hợp để tránh sinh ra các data mart thừa

Làm gì với dữ liệu xấu:

- Dữ liệu hàng ngày có thể gom được từ các nguồn khác nhau chưa chắc là tốt hoàn toàn sẽ được nhập kho dữ liệu.

- Xác định nguồn gốc dữ liệu xấu

ENTRY cân bằng

- Tìm thấy những dữ liệu xấu, thì 1 entry tương đương sẽ sửa lại nó.
- Phương pháp này chỉ hoạt động, nơi có một số lượng hữu hạn của dữ liệu được điều chỉnh
- Dữ liệu sai có thể được xác định

Thiết lập lại giá trị

- Trong trường hợp không thể được các dữ liệu không chính xác cho một entry cân bằng ,được thực hiện bằng cách "reset" các giá trị cho một tài khoản.

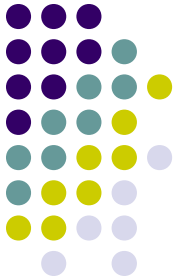
cách khác

- việc tìm kiếm bản ghi xấu
  và sau đó thay đổi các giá trị trong những bản ghi.

Nguyên nhân

- không xác định dc đúng vị trí của entry lỗi
- tính toàn vẹn của dữ liệu đã bị phá hủy

Vận tốc của Chuyển động dữ liệu



- Hệ thống này hoạt động nhanh hơn DW do có thể xử lí hàng loạt các query và xuất thông tin cùng 1 lúc bằng cách tạo ra các data mart phụ thuộc