ĐẠI HỌC KINH DOANH VÀ CÔNG NGHỆ HÀ NỘI

KHOA CNTT

Thu thập Tập dữ liệu lớn

CHƯƠNG 2: Thu thập Tập dữ liệu lớn

- **Đặc trưng dữ liệu lớn**
  - 3V
  - 5V
- **Thu thập dữ liệu**
  - Dữ liệu riêng
  - Dữ liệu ngoài
- **Làm sạch và tích hợp dữ liệu**
- **Các công cụ tích hợp dữ liệu**
  - Dell Boomi
  - Oracle Data Integrator
  - SAP Data Services
  - Snaplogic

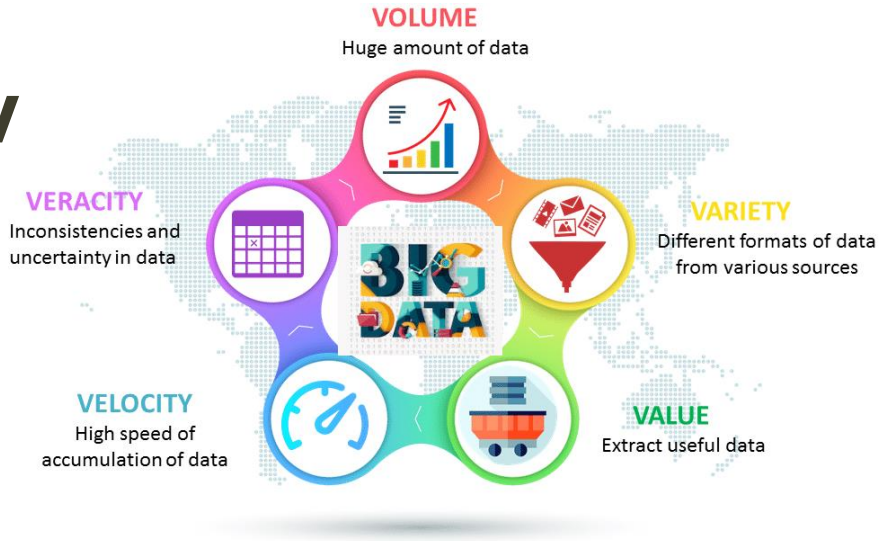
Đặc trưng Dữ liệu lớn

- Chúng ta có đang dùng 1 ứng dụng được tạo ra cách đây **10 năm**?
- Bạn có đang dùng các phần cứng được tạo ra cách đây **20 năm**?
- **Ban có đang dùng dữ liệu đã có cách đây 50 năm?** câu trả lời chắc chắn là có.

- **Big Data không chỉ là có rất nhiều dữ liệu**, nó là khái niệm cung cấp cơ hội để có cái nhìn sâu sắc (hiểu biết sâu) vào **dữ liệu hiện có** cũng như các hướng dẫn để thu thập và phân tích dữ liệu tương lai.

Đặc trưng Dữ liệu lớn

- **Đặc trưng cơ bản 3V**
  - Volume (Khối lượng)
  - Velocity (Tốc độ)
  - Variety (Đa dạng)



- **Đặc trưng bổ sung ~5V**
  - Veracity (Độ tin cậy/chính xác)
  - Value (Giá trị)

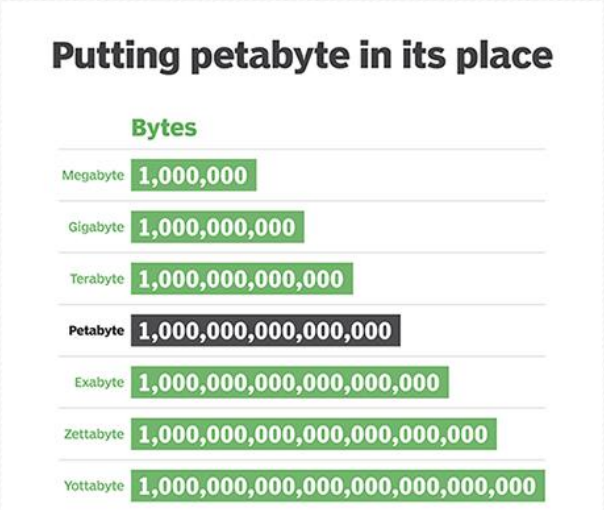
Đặc trưng Dữ liệu lớn

- **Volume (Dung lượng)** Số lượng dữ liệu được tạo ra và lưu trữ. Kích thước của dữ liệu xác định giá trị và tiềm năng insight (hiểu biết sâu) - và liệu nó có thể thực sự được coi là dữ liệu lớn hay không.
- **Velocity (Tốc độ)** Trong trường hợp này nghĩa là tốc độ các dữ liệu được tạo ra và xử lý để đáp ứng các nhu cầu và thách thức trên con đường tăng trưởng và phát triển.
- **Variety (Tính đa dạng)** Các dạng và kiểu của dữ liệu. Dữ liệu được thu thập từ nhiều nguồn khác nhau và các kiểu dữ liệu cũng có rất nhiều cấu trúc khác nhau.
- **Veracity (Độ tin cậy)** Chất lượng của dữ liệu thu được có thể khác nhau rất nhiều, ảnh hưởng đến sự phân tích chính xác.
- **Value (Giá trị)** Dữ liệu phải được xử lý bằng các công cụ tiên tiến (phân tích và thuật toán) để cho ra các thông tin có ý nghĩa.

Dung lượng:

- **Volume (Dung lượng - Khối lượng dữ liệu)** số lượng dữ liệu được tạo ra, lưu trữ.
- **Dung lượng** là sự tăng trưởng về mặt khối lượng. Dữ liệu trong các hệ thống thông tin luôn luôn và không ngừng tăng lên về mặt kích thước (khối lượng). Chúng ta có thể tìm thấy dữ liệu trong các định dạng video, music, image lớn trên các kênh truyền thông xã hội.

- **Khối lượng dữ liệu** là đặc điểm tiêu biểu nhất của dữ liệu lớn. Kích cỡ của Big data đang từng ngày tăng lên, và tính đến năm 2012 thì nó có thể nằm trong khoảng vài chục terabyte cho đến nhiều petabyte (1 petabyte = 1024 terabyte) chỉ cho một tập hợp dữ liệu.
- Dữ liệu truyền thống có thể lưu trữ trên các thiết bị đĩa mềm, đĩa cứng. Nhưng với **dữ liệu lớn** chúng ta sẽ sử dụng các công nghệ lưu trữ đặc thù hoặc phổ biến ngày nay là công nghệ "đám mây" mới đáp ứng khả năng lưu trữ được dữ liệu lớn.



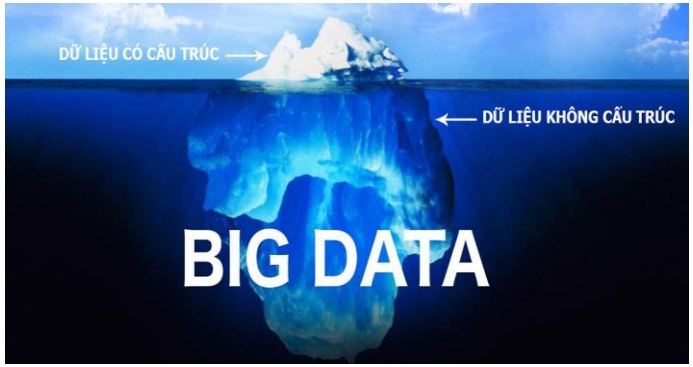
Tốc độ:

- **Velocity (Tốc độ)** là tốc độ xử lý các dữ liệu được tạo ra để đáp ứng các nhu cầu và thách thức trích xuất thông tin mang lại lợi ích cho tăng trưởng và phát triển.
- Bên cạnh sự tăng trưởng về khối lượng, thời gian cập nhật của dữ liệu cũng rút ngắn hơn ngày càng nhanh. Các ứng dụng phổ biến trên lĩnh vực Internet, Tài chính, Ngân hàng, Hàng không, Quân sự, Y tế – Sức khỏe như hiện nay phần lớn dữ liệu lớn được xử lý real-time. Công nghệ xử lý dữ liệu lớn ngày nay đã cho phép chúng ta xử lý tức thì trước khi chúng được lưu trữ vào cơ sở dữ liệu.

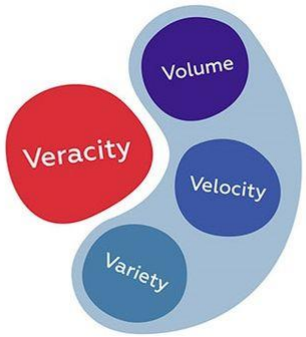
- **Tốc độ đáp ứng xử lý dữ liệu có thể hiểu theo 2 khía cạnh:**
  - Khối lượng dữ liệu gia tăng rất nhanh (mỗi giây có tới 72.9 triệu các yêu cầu truy cập tìm kiếm trên web bán hàng của Amazon);
  - Xử lý dữ liệu nhanh ở mức thời gian thực (real-time), có nghĩa dữ liệu được xử lý ngay tức thời ngay sau khi chúng phát sinh (tính đến bằng mili giây).

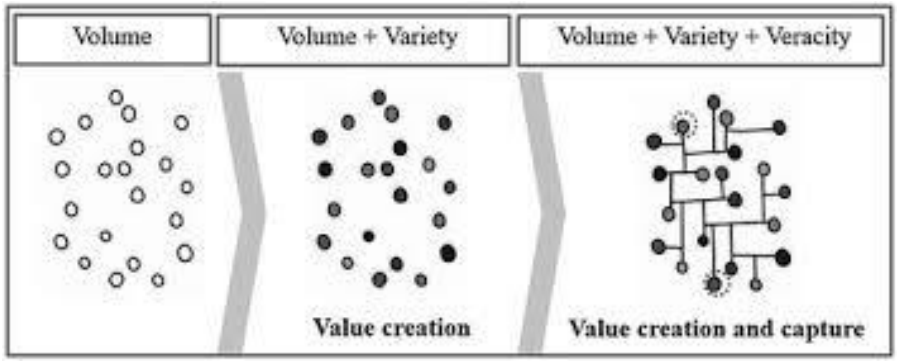
Tính đa dạng:

- **Variety (Tính đa dạng)** Các dạng và kiểu của dữ liệu. Dữ liệu được thu thập từ nhiều nguồn khác nhau và các kiểu dữ liệu cũng có rất nhiều cấu trúc khác nhau.
- **Variety** là sự phát triển về tính đa dạng của dữ liệu. Dữ liệu của một doanh nghiệp hay một hệ thống thông tin ngày nay không còn đơn giản chỉ có một hoặc một vài loại dữ liệu nữa, mà tính đa dạng của nó cũng đang ngày càng tăng lên làm cho tính phức tạp của dữ liệu ngày càng phức tạp hơn.
- Đối với dữ liệu truyền thống chúng ta hay nói đến dữ liệu có cấu trúc, thì ngày nay hơn 80% dữ liệu được sinh ra là phi cấu trúc (tài liệu, blog, hình ảnh, vi deo, bài hát, dữ liệu từ thiết bị cảm biến vật lý, thiết bị chăm sóc sức khỏe…).
- Big data cho phép liên kết và phân tích nhiều dạng dữ liệu khác nhau. Ví dụ, với các bình luận của một nhóm người dùng nào đó trên Facebook với thông tin video được chia sẻ từ Youtube và Twitter.



Độ tin cậy:

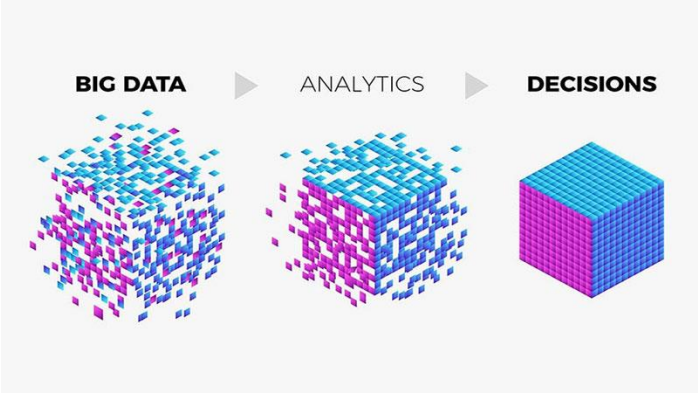




- Chất lượng của dữ liệu thu được có thể khác nhau rất nhiều, ảnh hưởng đến sự phân tích chính xác.
- **Veracity** là tính xác thực của dữ liệu. Với xu hướng mở rộng liên kết và sự gia tăng mạnh mẽ tính tương tác và chia sẻ của người dùng trực tuyến làm cho bức tranh xác định về độ tin cậy & chính xác của dữ liệu ngày một khó khăn hơn. Bài toán phân tích và loại bỏ dữ liệu thiếu chính xác và gây nhiễu đang là tính chất quan trọng của Big Data.
- **Độ tin cậy** chính là một trong những tính chất phức tạp nhất của Dữ liệu lớn - độ tin cậy/chính xác của dữ liệu.

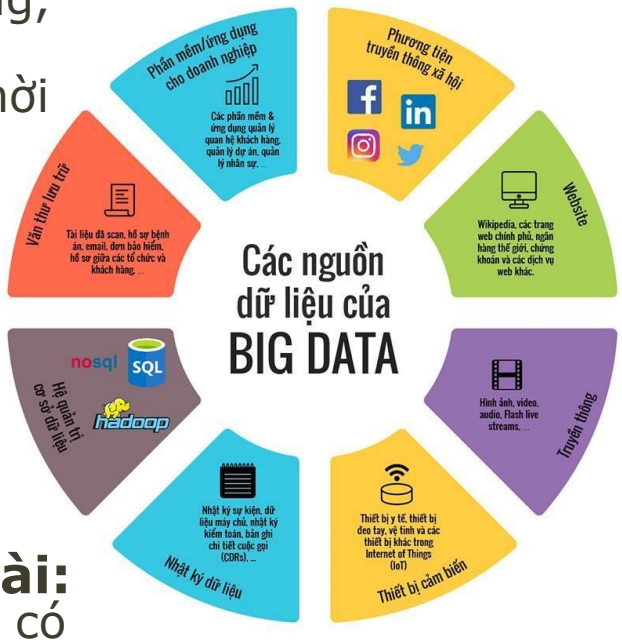
Giá trị:

- **Value (Giá trị)** Dữ liệu phải được xử lý bằng các công cụ tiên tiến (phân tích và thuật toán) để cho ra các thông tin có ý nghĩa.
- Giá trị là đặc điểm quan trọng nhất của dữ liệu lớn, vì khi bắt đầu triển khai xây dựng dữ liệu lớn thì việc đầu tiên chúng ta cần phải làm đó là xác định được giá trị của thông tin mang lại như thế nào, khi đó chúng ta mới có quyết định có nên triển khai dữ liệu lớn hay không.
- Nếu chúng ta có dữ liệu lớn mà chỉ nhận được 1% lợi ích từ nó, thì không nên đầu tư phát triển dữ liệu lớn. Kết quả dự báo chính xác thể hiện rõ nét nhất về giá trị của dữ liệu lớn mang lại.



Thu thập dữ liệu

- **Dữ liệu thu thập từ các nguồn**
  - Truyền thống: thông tin khách hàng, giao dịch…
  - Thu thập tự động qua cảm biến: thời tiết, nhật ký…
  - Mạng xã hội: comment trên facebook, twitter…
- **Thu thập dữ liệu riêng**
  - Dữ liệu giao dịch
  - Dữ liệu thí nghiệm
  - Dữ liệu do người dùng tạo ra
  - Dữ liệu tổng hợp
- **Tiếp cận nguồn dữ liệu bên ngoài:** thứ ta tìm kiếm là các dữ liệu phù hợp, có thể trả lời và hỗ trợ các quyết định



Nguồn dữ liệu bên ngoài

Có nhiều phương án khác nhau giúp tiếp cận các dữ liệu bên ngoài và hiện đang gia tăng hàng ngày. Dữ liệu thu thập từ các nguồn:

- Dữ liệu chính phủ Việt Nam: www.Data.gov.vn
- Dữ liệu Tông cục thống kê Việt nam: www.gso.gov.vn
- Dữ liệu chính phủ Mỹ: www.data.gov
- Dữ liệu Quốc gia Anh: www.data.gov.uk
- Cổng dữ liệu mở Liên minh Châu Âu: https://data.europa.eu/en
- Thống kê tỷ lệ tìm kiếm Google: ww.google.com/trends/explore
- Dữ liệu bản đồ tương tác, hình ảnh vệ tinh, Street view trên Google Maps www.maps.google.com

Làm sạch và tối ưu hóa chất lượng dữ liệu

Làm sạch dữ liệu là quá trình biến đổi với những đối tượng dữ liệu không nhất quán, bị thiếu hụt, gây nhiễu… qua đó hệ thống sẽ có được dữ liệu sạch để có thể phân tích, đánh giá, dự đoán giá trị dữ liệu một cách chính xác.

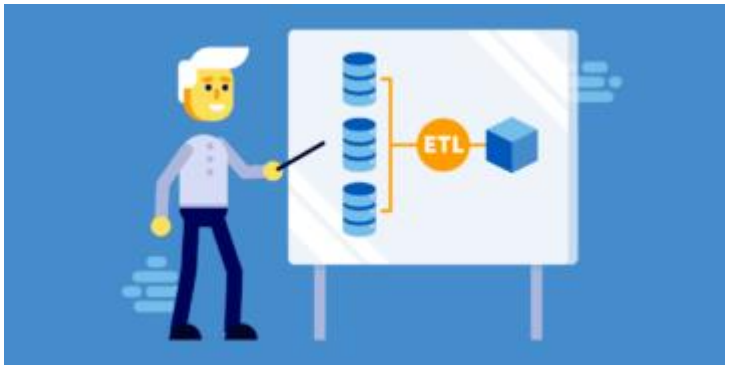
Quy trình làm sạch dữ liệu bao gồm các bước:

- *Xác định và xóa các tập dữ liệu không liên quan và trùng lặp.*
- *Sửa lỗi trong cấu trúc dữ liệu.*
- *Đưa ra các quy tắc làm sạch dữ liệu trong toàn tổ chức.*
- *Đầu tư vào các công cụ cho phép làm sạch dữ liệu trong thời gian thực.*

→ Đảm bảo dữ liệu phải luôn đáp ứng đầy đủ tính hoàn chỉnh, độc nhất, chính xác, tính nhất quán và hiệu lực.

Tích hợp dữ liệu – Data Integration

Tích hợp dữ liệu là quá trình kết hợp các dữ liệu không đồng nhất từ nhiều nguồn khác nhau tạo ra một lược đồ duy nhất từ đó có thể truy vấn, cung cấp cho hệ thống một cái nhìn tổng quan về các dữ liệu đó,



Tích hợp dữ liệu được sử dụng với tần số ngày càng nhiều khi mà khối lượng và nhu cầu chia sẻ dữ liệu hiện nay rất lớn. Để đảm bảo việc trao đổi dữ liệu trong hệ thống được hiệu quả hoặc xử lý các công việc tiếp theo theo các luồng công việc định trước.

Tích hợp dữ liệu cần thiết để đạt được giá trị gia tăng từ những tài nguyên, thành phần phần đang tồn tại và lưu từ phân tán

Tích hợp dữ liệu – Data Integration

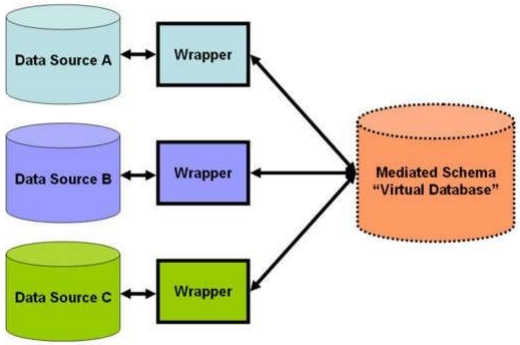
Có hai phương pháp chủ yếu để tích hợp dữ liệu:

- **Ghép nối chặt chẽ (Tight Coupling): kho dữ liệu vật lý**

Phương pháp ghép nối chặt chẽ thường được thực hiện thông qua kho dữ liệu, dữ liệu được lấy từ nhiều nguồn khác nhau đưa vào một vị trí vật lý duy nhất thông qua quá trình ETL (Extraction, Transformation, Loading). Lớp ETL giúp ánh xạ dữ liệu từ các nguồn để cung cấp một kho dữ liệu thống nhất , cung cấp một giao diện đồng nhất để truy vấn dữ liệu. Cách tiếp cận này được gọi là ghép nối chặt chẽ vì trong cách tiếp cận này dữ liệu được kết hợp chặt chẽ với kho lưu trữ vật lý tại thời điểm truy vấn.

- **Ghép nối lỏng lẻo (Loose Coupling): lược đồ trung gian ảo**

Ở đây một lược đồ trung gian ảo cung cấp một giao diện nhận truy vấn từ người dùng, biến đổi nó theo cách mà cơ sở dữ liệu có thể hiểu và gửi truy vấn trực tiếp tới cơ sở dữ liệu nguồn để thu được kết quả. Trong phương pháp này, dữ liệu chỉ nằm trong cơ sở dữ liệu nguồn thực tế.



Tích hợp dữ liệu

	GHÉP NỐI CHẶT CHẼ	GHÉP NỐI LỎNG LẺO
Ưu điểm	– Độc lập (phụ thuộc ít hơn vào hệ thống nguồn vì dữ liệu được sao chép về mặt vật lý) – Xử lý truy vấn nhanh hơn – Xử lý truy vấn phức tạp – Tóm tắt dữ liệu nâng cao và có thể lưu trữ – Xử lý dữ liệu lớn	– Làm mới dữ liệu (độ trễ thấp – gần như thời gian thực) – Nhanh hơn (khi có hệ thống nguồn mới hoặc hệ thống nguồn hiện tại thay đổi thì chỉ bộ điều hợp tương ứng mới được tạo hoặc thay đổi, phần lớn không ảnh hưởng đến các phần khác của hệ thống) – Chi phí thấp
Nhược điểm	– Độ trễ (vì dữ liệu cần được tải bằng ETL) – Chi phí cao	– Phản hồi truy vấn chậm hơn (do vấn đề về mạng / băng thông, tải dữ liệu trên hệ thống nguồn,…) – Phụ thuộc vào các nguồn dữ liệu

Công cụ tích hợp dữ liệu

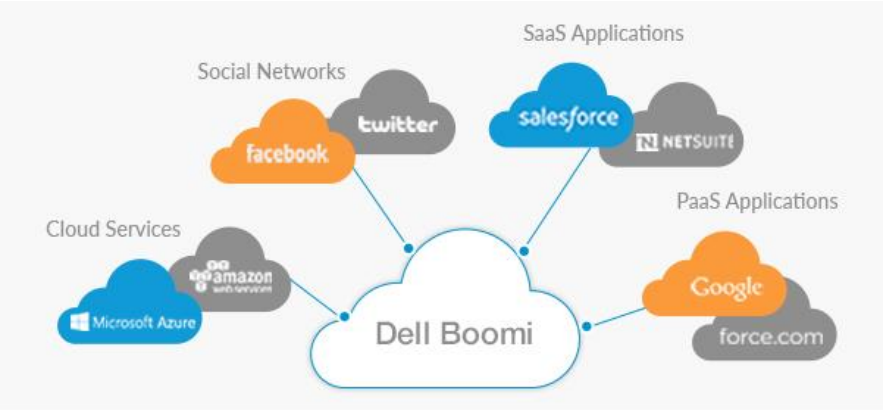
Có nhiều công cụ tích hợp dữ liệu có sẵn trên thị trường được phát triển cập nhật liên tục. Tất cả đều chạy đua để theo kịp sự gia tăng nhanh chóng của dữ liệu, điện toán đám mây và dữ liệu lớn.

Dell Boomi

- Dell Boomi là nền tảng tích hợp đám mây đa người thuê, kết nối dữ liệu và ứng dụng tại chỗ với đám mây, hỗ trợ thiết kế và quản lý IoT, API và quản lý dữ liệu tổng thể. Thị trường mục tiêu của Dell Boomi được thiết lập từ thị trường trung bình đến các doanh nghiệp lớn.
- Boomi hỗ trợ khách hàng và đối tác khai phá, quản lý, sắp xếp dữ liệu nhanh chóng và dễ dàng, đồng thời kết nối các ứng dụng, quy trình để giúp có kết quả xử lý nhanh hơn.

Tính năng, đặc điểm:

- Giao diện trực quan dễ sư dụng
- Cho phép thiết kế các quy trình tích hợp
- Tự động hóa quy trình làm việc
- Tốc độ xử lý nhanh
- Cập nhật tích hợp tự động
- Hỗ trợ cho một loạt các giao thức IoT
- Giám sát hoạt động và theo dõi sự kiện



SAP Data Services

- **SAP Data Services** cung cấp giải pháp theo mức độ từng doanh nghiệp để tích hợp dữ liệu, chuyển đổi chất lượng dữ liệu, định hình dữ liệu và xử lý dữ liệu văn bản từ nhiều nguồn dữ liệu khác nhau vào cơ sở dữ liệu đích hoặc kho dữ liệu. Nó có thể được sử dụng độc lập hoặc với các sản phẩm khác của **SAP ERP (Enterprise Resource Planning)**

- **Tính năng, đặc điểm:**
  - SAP Data Service là công cụ giúp khám phá, làm sạch, tăng cường, tích hợp và quản lý dữ liệu từ các nguồn SAP và Không phải của SAP, bao gồm hệ cơ sở dữ liệu quan hệ, ứng dụng doanh nghiệp, tệp và dữ liệu lớn như là Hadoop và hệ cơ sở dữ liệu NoQuery
  - Cập nhật quan trọng trong phiên bản gần đây, 4.2 bao gồm hỗ trợ SAP HANA hỗ trợ được cải thiện ngoài khả năng kết nối tốt hơn với các nguồn dữ liệu lớn và xử lý văn bản và XML.

ODI - Oracle Data Integrator

- **Oracle Data integrator** 12c là nền tảng tích hợp dữ liệu tốt dành cho các tổ chức sử dụng các hệ thống và ứng dụng khác của Oracle, mong muốn tích hợp dư liệu gần gũi với các hệ thống này.
- Nền tảng kết hợp với cơ sở dữ liệu Oracle, Oracle GoldenGate, Oracle Fusion Middleware, Oracle Big Data Appliance và Exadata.

- **Tính năng, đặc điểm:**

• Giao diện người dùng dựa trên luồng khai báo

• Hỗ trợ nhiều mục tiêu

• Cải tiến hiệu suất thời gian chạy

• Hỗ trợ tất cả các RDBMS bao gồm tất cả các nền tảng dữ liệu hàng đầu như:

Oracle, Exadata, Teradata, IBM DB2, Netezza, Sybase IQ, MSSserver và nhiều công nghệ khác như tệp phẳng, ERP, LDAP, XML

ORACLE®
DATA INTEGRATOR

Snaplogic

- **SnapLogic** là một dịch vụ nền tảng tích hợp cho phép kết nối nhanh hơn và cập nhật kịp sự phát triển không ngừng của các hệ thống dữ liệu. Công cụ hỗ trợ trong việc áp dụng các ứng dụng Đám mây

- **Tính năng, đặc điểm:**

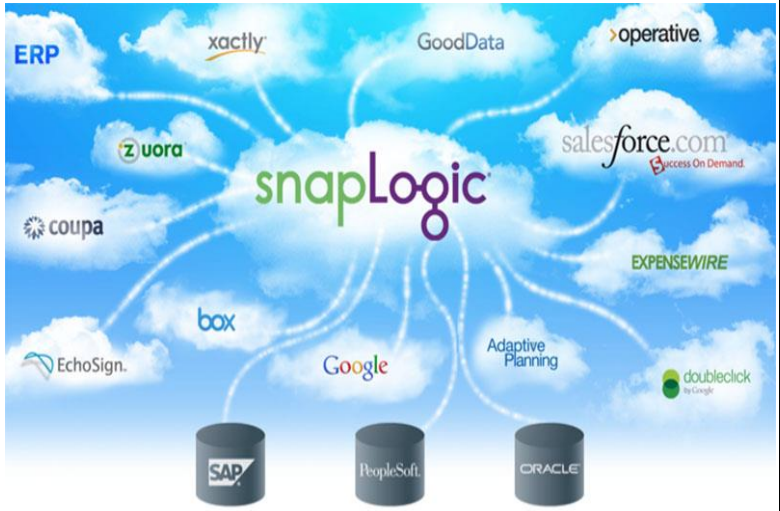
• Cung cấp giao diện trực quan mà không cần mã hóa

• Cung cấp công cụ để quản lý luồng dữ liệu trong suốt vòng đời của nó từ việc tạo và lưu trữ ban đầu đến loại bỏ

• Hỗ trợ tích hợp dựa trên sự kiện hoặc giao dịch chống lại các thay đổi theo thời gian thực.

• Hỗ trợ các trình kết nối khác nhau trên SaaS, Enterprise, Big Data, Mainframe, Files

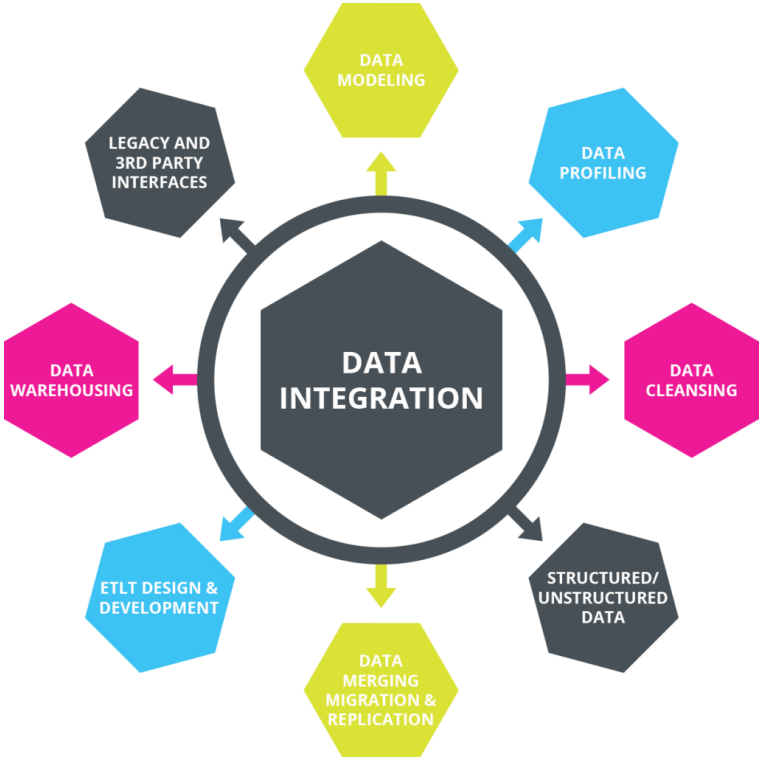
• Cung cấp tích hợp cho các nguồn Dữ liệu lớn như Hadoop và các nguồn NoQuery khác



Bài tập nhóm

Thuyết trình các chức năng cơ bản được cung cấp trên nền tảng:

- Dell Boomi
- SQL Server Integration Services (SSIS)
- Oracle Data Integrator
- SAP Data Services
- Talend Platform for Big Data Integration
- Pentaho Platform




Cảm ơn các bạn đã
lắng nghe!