ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

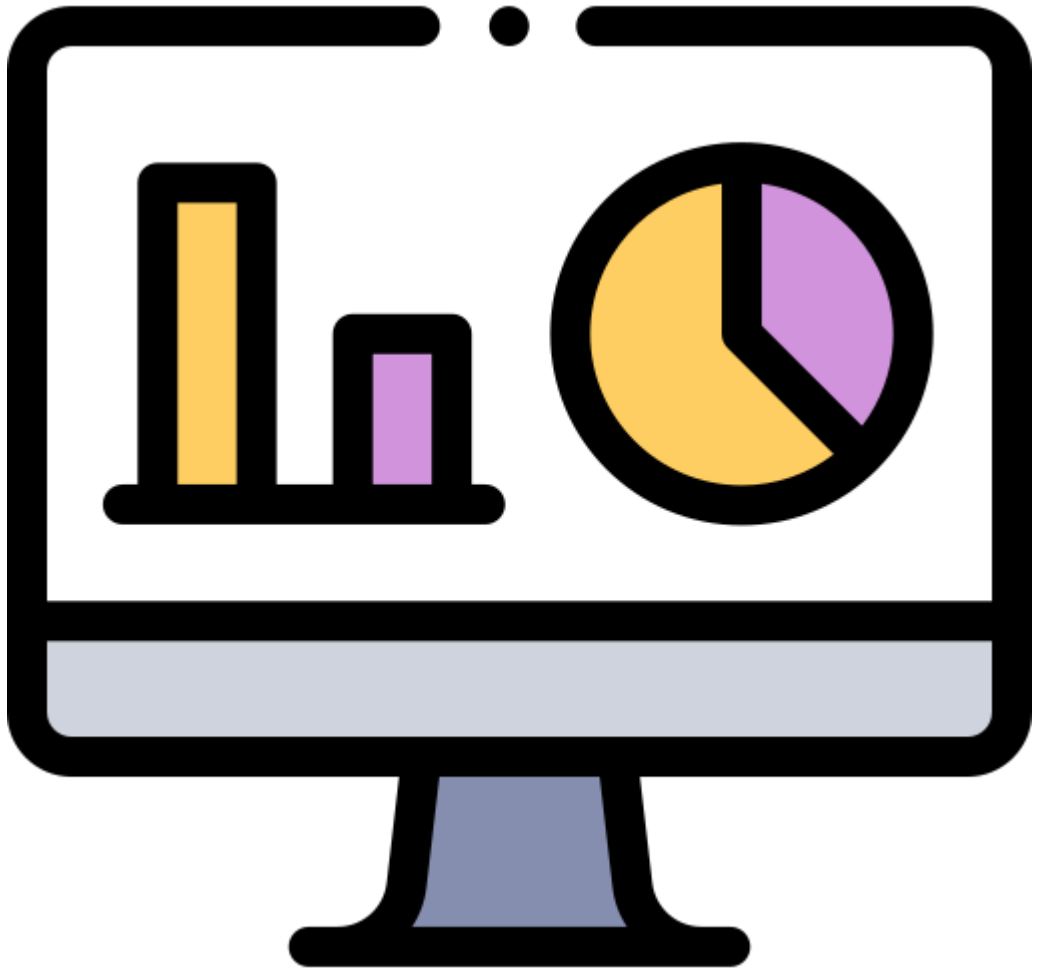
CHƯƠNG 5
Phân tích dữ liệu cho doanh nghiệp

Biên soạn: ThS. Nguyễn Thị Anh Thư

Nội dung

1. Giới thiệu
2. Kiến trúc BI
3. BI trong kiến trúc dữ liệu lớn
4. Các thành phần chính của BI
5. Cung cấp thông tin
6. Tổng kết

1. Giới thiệu

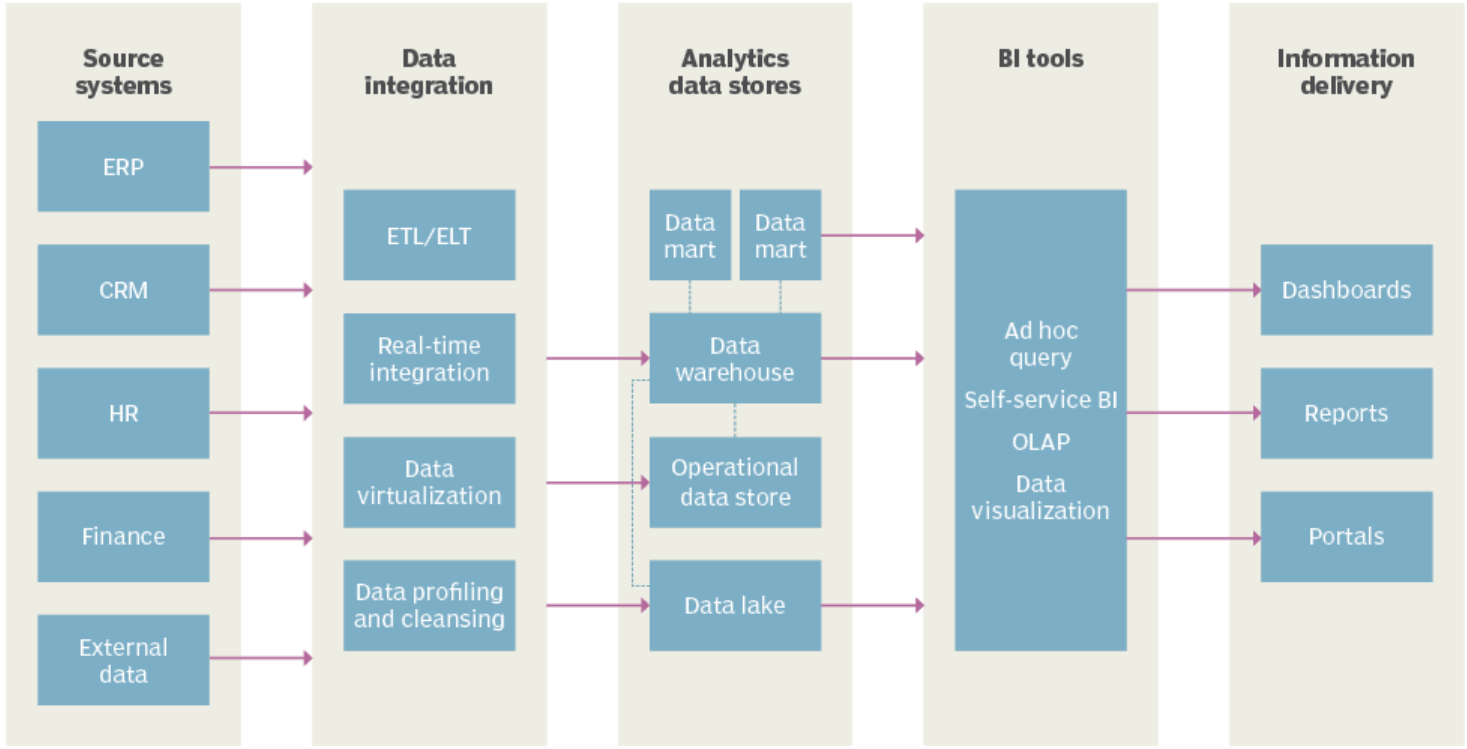


Phân tích dữ liệu cho doanh nghiệp – Business Intelligence (BI) là một tập hợp các công nghệ, quy trình và thực tiễn được sử dụng để **chuyển đổi dữ liệu thô thành thông tin hữu ích**, giúp doanh nghiệp đưa ra quyết định sáng suốt hơn và cải thiện hiệu quả hoạt động.

Lợi ích của BI:

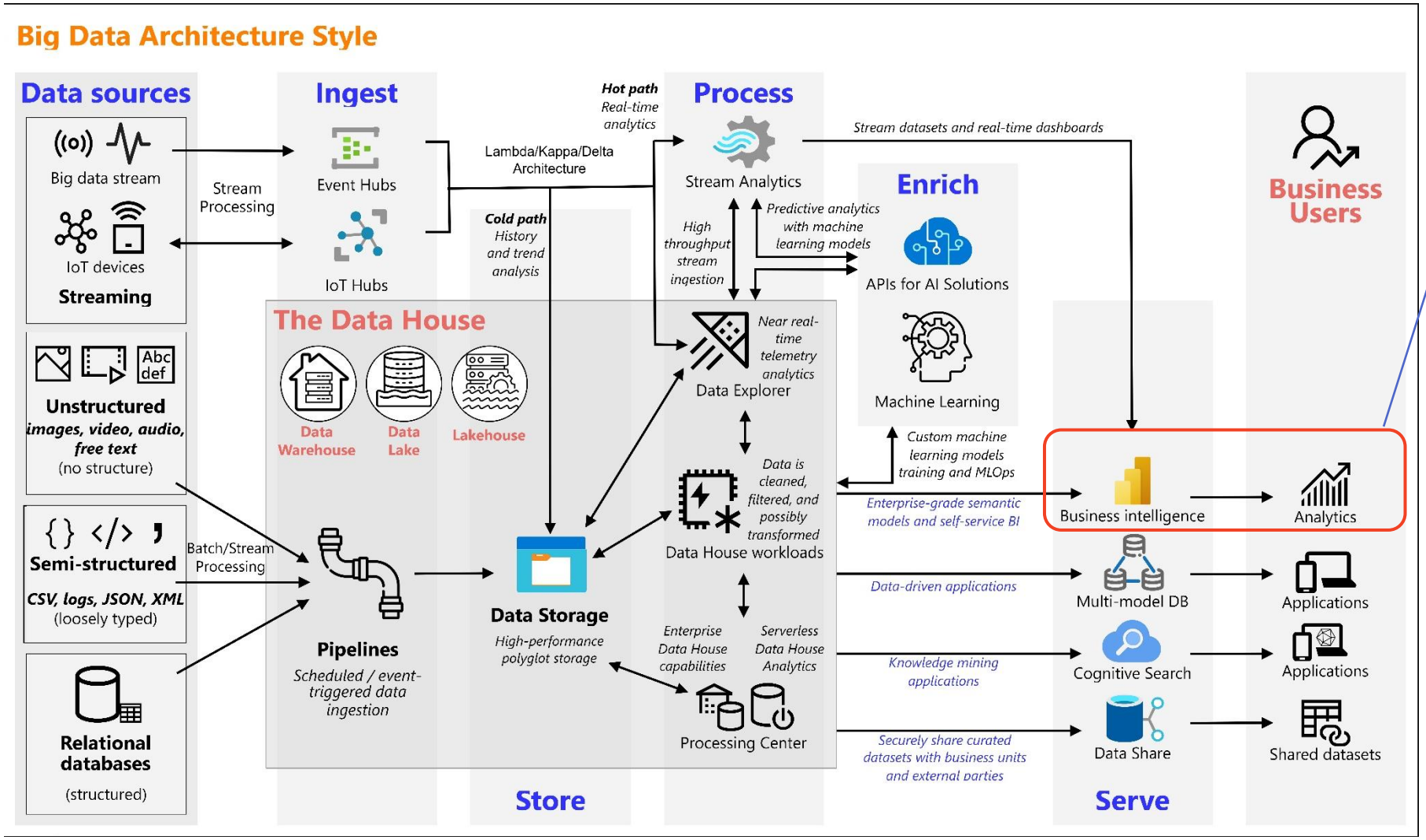
- Cải thiện hiệu quả hoạt động
- Tăng doanh thu
- Giảm chi phí
- Tăng lợi thế cạnh tranh

2. Kiến trúc BI



- **Source systems**: Dữ liệu nội bộ và bên ngoài.
- **Data integration and cleansing tools**: Các công cụ khám phá, làm sạch và chuyển đổi dữ liệu.
- **Analytics data stores**: Thường là dữ liệu có cấu trúc để truy vấn và phân tích.
- **BI and data visualization tools**: Phân tích dữ liệu và trình bày thông tin.
- **Dashboards, portals and reports**: Hiển thị kết quả.

3. BI trong kiến trúc dữ liệu lớn



- **Dashboards**: Gồm các biểu đồ, đồ thị và các hình ảnh trực quan dữ liệu. Tương tác với dữ liệu (*lọc dữ liệu theo thời gian, khu vực hoặc sản phẩm*)
- **Reports**: Gồm các tính năng Trực quan hóa dữ liệu, Lọc dữ liệu, Sắp xếp dữ liệu, Xuất dữ liệu.
- **Portals**: Bao gồm dashboards, báo cáo, phân tích dữ liệu và kho dữ liệu.

4. Các thành phần chính của BI

- Thu thập dữ liệu
- Lưu trữ dữ liệu
- Xử lý dữ liệu
- Phân tích dữ liệu
- Trực quan hóa dữ liệu

Thu thập dữ liệu

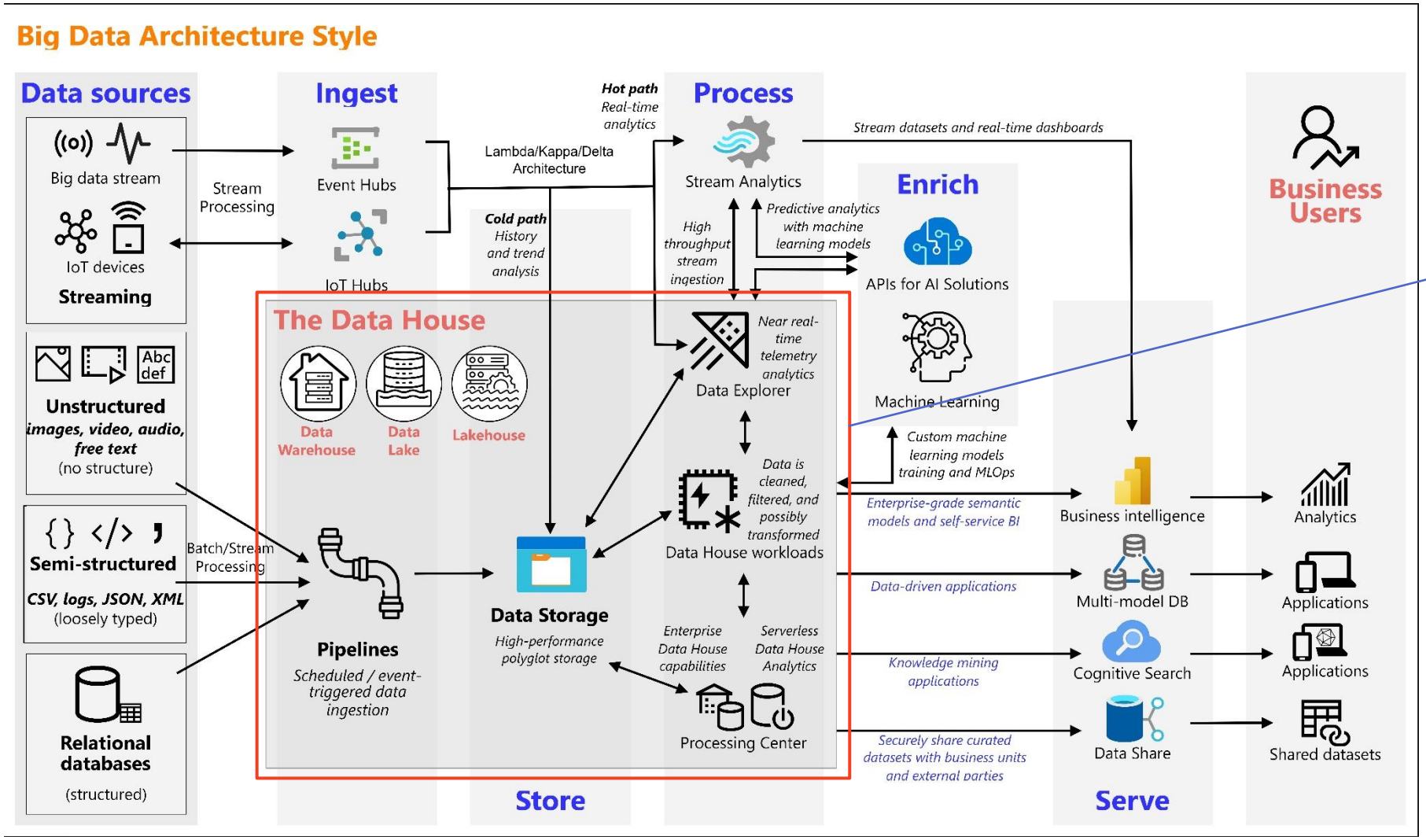
Dữ liệu nội bộ

- **ERP**: Hệ thống hoạch định nguồn lực doanh nghiệp (Tài chính, Kế toán, Nhân sự, Bán hàng và Tiếp thị, …)
- **CRM**: Hệ thống quản lý quan hệ khách hàng (Thông tin khách hàng, Lịch sử mua hàng, Tương tác với khách hàng, Phản hồi của khách hàng, …)
- **HR**: Tập hợp thông tin liên quan đến nhân viên của một tổ chức.
- **Finance**: Tập hợp thông tin về tình hình tài chính của một tổ chức.

Dữ liệu bên ngoài

- Dữ liệu thị trường
- Dữ liệu mạng xã hội
- Dữ liệu cảm biến, …

Lưu trữ dữ liệu



Dữ liệu được lưu trữ trong kho dữ liệu để truy cập và phân tích dễ dàng.

Xử lý dữ liệu

Dữ liệu thô được làm sạch, sắp xếp và chuẩn hóa trước khi có thể được phân tích.

- Làm sạch dữ liệu cơ bản bằng cách xử lý các giá trị thiếu, nhiễu và không nhất quán.
- **Xử lý nhiễu**
  - df.describe(): Thống kê tóm tắt về dữ liệu, giúp xác định các giá trị ngoại lai.
  - df.boxplot(): Tạo biểu đồ boxplot để trực quan hóa các giá trị ngoại lai.
  - df.replace(): Thay thế các giá trị ngoại lai bằng giá trị khác.
  - StandardScaler(): Thư viện từ scikit-learn để chuẩn hóa dữ liệu, giúp giảm thiểu ảnh hưởng của nhiễu.

Xử lý dữ liệu

Dữ liệu thô được làm sạch, sắp xếp và chuẩn hóa trước khi có thể được phân tích.

- Làm sạch dữ liệu cơ bản bằng cách xử lý các giá trị thiếu, nhiễu và không nhất quán.

- **Xử lý dữ liệu không nhất quán**
  - df.dtypes: Kiểm tra kiểu dữ liệu của các cột.
  - df.astype(): Chuyển đổi kiểu dữ liệu của các cột.
  - df.replace(): Thay thế các giá trị không nhất quán bằng giá trị hợp lệ.
  - dict(): Tạo dictionary để ánh xạ các giá trị không nhất quán thành giá trị hợp lệ.

Xử lý dữ liệu

Dữ liệu thô được làm sạch, sắp xếp và chuẩn hóa trước khi có thể được phân tích.

- Chuyển đổi dữ liệu sang dạng phù hợp cho việc phân tích, ví dụ như mã hóa dữ liệu categorical.

Pandas	NumPy
• df.fillna(value): Thay thế giá trị thiếu bằng giá trị được chỉ định. • df.replace(to_replace, value): Thay thế giá trị cụ thể bằng giá trị khác. • df.astype(dtype): Chuyển đổi kiểu dữ liệu của cột sang kiểu dữ liệu mong muốn. • df.convert_dtypes(): Tự động chuyển đổi kiểu dữ liệu của cột sang kiểu phù hợp nhất. • df.dropna(axis=0, thresh=n): Loại bỏ các hàng có giá trị thiếu (axis=0) hoặc giữ lại các hàng có ít nhất n giá trị không thiếu (thresh=n).	• np.nan_to_num(x): Chuyển đổi giá trị NaN thành số. • np.where(condition, x, y): Thay thế các phần tử trong mảng dựa trên điều kiện. • np.array(data, dtype): Chuyển đổi dữ liệu sang dạng mảng NumPy với kiểu dữ liệu mong muốn. • np.reshape(array, shape): Thay đổi hình dạng của mảng NumPy.

Xử lý dữ liệu

Dữ liệu thô được làm sạch, sắp xếp và chuẩn hóa trước khi có thể được phân tích.

- Kết hợp các tập dữ liệu từ nhiều nguồn khác nhau.
  - **Pandas**: Cung cấp các hàm như *concat()* và *merge()* để kết hợp các DataFrame theo hàng, cột hoặc khóa.
  - **Dask**: Cung cấp các API tương tự như Pandas, bao gồm *concat()* và *merge()* để kết hợp các DataFrame.
  - **NumPy**: Cung cấp các hàm như *concatenate()* và *stack()* để kết hợp các mảng.

Phân tích dữ liệu

Phân tích thống kê

- **Mô tả thống kê**: Bao gồm các chỉ số như trung bình, trung vị, độ lệch chuẩn, v.v., giúp mô tả đặc điểm cơ bản của dữ liệu.
- ***Thử nghiệm thống kê***: Giúp kiểm định giả thuyết và xác định mức độ tin cậy của kết quả phân tích.
- **Hồi quy tuyến tính**: Giúp dự đoán giá trị của một biến dựa trên các biến khác.
- **Phân tích nhóm**: Giúp phân chia dữ liệu thành các nhóm khác nhau dựa trên đặc điểm chung.
- ***Phân tích chuỗi thời gian***: Giúp dự đoán xu hướng tương lai dựa trên dữ liệu quá khứ.

Phân tích dữ liệu

Phân tích thống kê – *Thử nghiệm thống kê*

Phương pháp kiểm định giả thuyết t-test là một công cụ thống kê giúp *so sánh giá trị trung bình (mean) của hai nhóm dữ liệu*.

Cách thức hoạt động:

- *Đặt giả thuyết*:
  - Giả thuyết null (H0): Không có sự khác biệt giữa hai nhóm.
  - Giả thuyết thay thế (H1): Có sự khác biệt giữa hai nhóm.
- *Tính toán thống kê t*: Sử dụng công thức t-test để tính toán thống kê t dựa trên dữ liệu của hai nhóm.
- *Xác định giá trị p*: Sử dụng bảng phân phối t với bậc tự do phù hợp để xác định giá trị p.
- *Kết luận*: So sánh giá trị p với mức ý nghĩa (α) đã chọn (thường là 0.05).
  - Nếu $p < \alpha$, bác bỏ H0 và chấp nhận H1.
  - Nếu $p \geq \alpha$, không đủ bằng chứng để bác bỏ H0.

Phân tích dữ liệu

Phân tích thống kê – *Thử nghiệm thống kê*

Phương pháp kiểm định giả thuyết t-test là một công cụ thống kê giúp *so sánh giá trị trung bình (mean) của hai nhóm dữ liệu*.

- *Xác định sự khác biệt giữa các nhóm*: So sánh giá trị trung bình của hai hoặc nhiều nhóm dữ liệu để xác định liệu có sự khác biệt thống kê đáng kể hay không.
- *Xác định mối liên hệ giữa các biến*: Sử dụng t-test để kiểm tra mối liên hệ giữa hai biến.
  - Ví dụ như mối liên hệ giữa tuổi tác và thu nhập.
    - Hai nhóm khách hàng với mức tuổi tác khác nhau.
    - Kiểm tra xem liệu có sự khác biệt về thu nhập giữa hai nhóm này hay không.

Phân tích dữ liệu

Phân tích thống kê – *Phân tích chuỗi thời gian*

Hiểu rõ hơn về sự thay đổi của dữ liệu theo thời gian và dự đoán xu hướng tương lai.

- **Chuỗi thời gian** là tập hợp các giá trị được ghi lại theo thời gian. Ví dụ như doanh thu theo tháng, giá cổ phiếu theo ngày, v.v.
- Phân tích chuỗi thời gian tập trung vào việc *xác định các mẫu, xu hướng và mùa vụ* trong dữ liệu, từ đó dự đoán giá trị tương lai.
  - **Phân tích xu hướng**: Sử dụng các ***phương pháp như trung bình di động***, ARIMA, Holt-Winters để xác định xu hướng.

Phân tích dữ liệu

Phân tích thống kê – *Phân tích chuỗi thời gian*

Phương pháp trung bình di động (Moving Average - MA) tính toán trung bình của một số giá trị gần đây nhất trong chuỗi thời gian để dự đoán giá trị tiếp theo.

Cách thức hoạt động:

- **Chọn chu kỳ**: Xác định số lượng giá trị sẽ được sử dụng để tính trung bình. Chu kỳ phổ biến là 5, 10, 20, 50, 100, v.v.
- **Tính trung bình**: Tính trung bình của số lượng giá trị được chọn trong chu kỳ.
- **Dự báo**: Sử dụng giá trị trung bình được tính toán làm dự đoán cho giá trị tiếp theo.

Ví dụ: Sử dụng giá trị trung bình được tính toán từ dữ liệu giá đóng cửa của một cổ phiếu trong 5 ngày qua làm dự đoán cho giá cổ phiếu ngày hôm nay.

Phân tích dữ liệu

Khai phá tri thức

- Xác định các mẫu, xu hướng và mối quan hệ ẩn trong dữ liệu.
- Sử dụng các kỹ thuật như khai phá luật kết hợp để khám phá các quy tắc ẩn trong dữ liệu.
- Tóm tắt và diễn giải các kết quả phân tích để dễ dàng hiểu và áp dụng.
- **Khai phá luật kết hợp (Association Rule Mining)**
  - ***apriori***(transactions, min_support, min_confidence): Hàm khai phá luật kết hợp sử dụng thuật toán Apriori.
  - *fp_growth*(transactions, min_support): Hàm khai phá luật kết hợp sử dụng thuật toán FP-Growth.
  - *eclat*(transactions, min_support): Hàm khai phá luật kết hợp sử dụng thuật toán Eclat.

Phân tích dữ liệu

Khai phá tri thức

Hàm khai phá luật kết hợp sử dụng thuật toán Apriori

- Tìm kiếm các mối liên hệ tiềm ẩn giữa các tập hợp mục (itemset) trong cơ sở dữ liệu giao dịch.
- Thuật toán hoạt động dựa trên nguyên tắc “tính lặp” và “hỗ trợ” để xác định các tập hợp mục thường xuyên xuất hiện cùng nhau và có mức độ hỗ trợ cao.

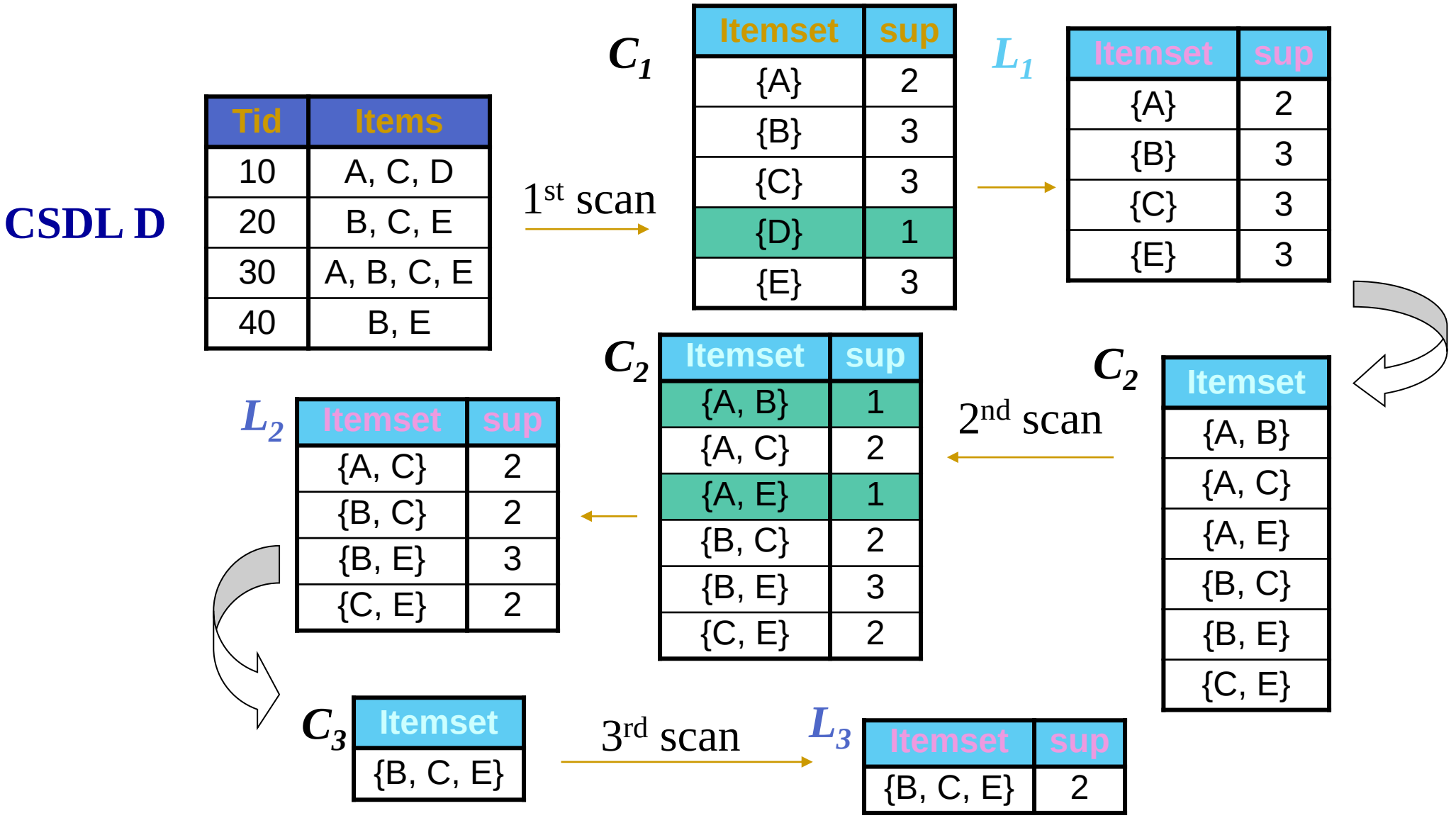
Phân tích dữ liệu

Khai phá tri thức

Hàm khai phá luật kết hợp sử dụng thuật toán Apriori

- Cách thức hoạt động:
  - *Đếm số lần xuất hiện của từng mục*: Thuật toán Apriori đếm số lần xuất hiện của từng mục trong tập dữ liệu giao dịch. Các mục có số lần xuất hiện cao hơn một ngưỡng nhất định sẽ được chọn làm tập hợp mục L1.
  - *Tạo tập hợp mục ứng viên*: Từ L1, thuật toán Apriori tạo ra các tập hợp mục ứng viên L2 bằng cách kết hợp từng cặp mục trong L1.
  - *Kiểm tra*: L2 được quét qua để kiểm tra xem mỗi tập hợp mục ứng viên có đạt mức độ hỗ trợ tối thiểu (min_support) hay không. Các tập hợp mục ứng viên không đạt yêu cầu sẽ bị loại bỏ.
  - *Lặp lại*: Quá trình tạo tập hợp mục ứng viên và kiểm tra hỗ trợ được lặp lại cho đến khi không còn tập hợp mục ứng viên nào được tạo ra.

Thuật toán Apriori



Phân tích dữ liệu

Khai phá tri thức

Hàm khai phá luật kết hợp sử dụng thuật toán Apriori

- Giả sử chúng ta có tập dữ liệu giao dịch của một cửa hàng bán lẻ với các mục sau:
  - A: Sữa
  - B: Bánh mì
  - C: Trứng
  - D: Nước trái cây
- Với min_support là 50%:

Phân tích dữ liệu

Khai phá tri thức

Hàm khai phá luật kết hợp sử dụng thuật toán Apriori

- Đếm số lần xuất hiện:
  - A: 60%, B: 70%, C: 50%, D: 40%
- Tạo L1: L1 = {A, B, C}
- Tạo L2: L2 = {AB, AC, BC}
- Kiểm tra support :
  - AB: 60%
  - AC: 55%
  - BC: 50%

Phân tích dữ liệu

Khai phá tri thức

Hàm khai phá luật kết hợp sử dụng thuật toán Apriori

Kết quả:

Luật kết hợp: **conf (A ⇒ (S-A)) = supp (S) / supp (A)**

- A => B (60%)
- B => A (60%)

Giải thích:

- Luật kết hợp A => B (60%) có nghĩa là 60% khách hàng mua sữa cũng mua bánh mì.
- Luật kết hợp B => A (60%) có nghĩa là 60% khách hàng mua bánh mì cũng mua sữa.

Trực quan hóa dữ liệu

Dữ liệu được trình bày dưới dạng biểu đồ, đồ thị và bảng để dễ dàng hiểu và chia sẻ.

- Sử dụng các biểu đồ như scatter plot, heatmap, line plot, v.v. để thể hiện mối quan hệ giữa các biến.
  - **Matplotlib**: Phù hợp cho việc vẽ đồ thị đơn giản và tùy chỉnh cao.
  - **Seaborn**: Phù hợp cho việc vẽ đồ thị đẹp mắt và dễ sử dụng.
  - **Plotly**: Phù hợp cho việc vẽ đồ thị tương tác và chia sẻ trực tuyến.
  - **Bokeh**: Phù hợp cho việc vẽ đồ thị tương tác và tích hợp với các ứng dụng web.
- Ví dụ: Matplotlib
  - Scatter plot: plt.scatter(x, y)
  - Heatmap: plt.imshow(data)
  - Line plot: plt.plot(x, y)

Trực quan hóa dữ liệu

Dữ liệu được trình bày dưới dạng biểu đồ, đồ thị và bảng để dễ dàng hiểu và chia sẻ.

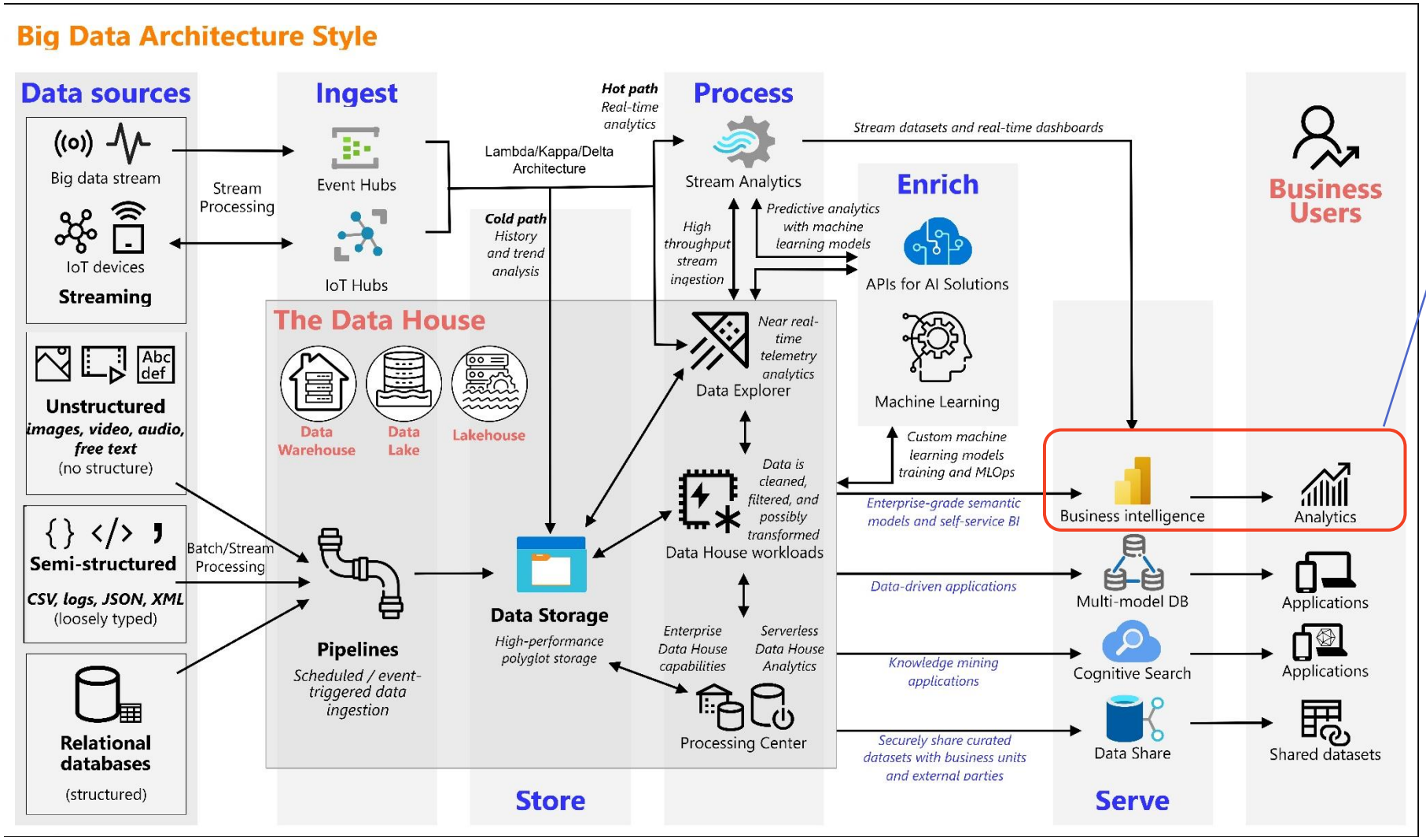
- Tạo các dashboard trực quan để theo dõi các chỉ số quan trọng và khám phá các mẫu dữ liệu theo thời gian.
  - **Mức độ dễ sử dụng**: Streamlit là lựa chọn tốt nhất cho người mới bắt đầu.
  - **Tính linh hoạt**: Plotly và Dash cung cấp nhiều tính năng và khả năng tùy chỉnh.
  - **Khả năng tương tác**: Bokeh và Streamlit cung cấp khả năng tương tác thời gian thực.
  - **Loại biểu đồ**: Matplotlib hỗ trợ nhiều loại biểu đồ khác nhau.
- Ví dụ: streamlit
  - st.line_chart(data): Tạo biểu đồ
  - st.table(data): Tạo bảng

Trực quan hóa dữ liệu

Dữ liệu được trình bày dưới dạng biểu đồ, đồ thị và bảng để dễ dàng hiểu và chia sẻ.

- Sử dụng các kỹ thuật như dimensionality reduction để giảm bớt số lượng biến và dễ dàng trực quan hóa dữ liệu.
  - Hiểu rõ các phương pháp giảm chiều trước khi sử dụng.
  - Lựa chọn phương pháp phù hợp với dữ liệu và mục tiêu.
  - Thử nghiệm và so sánh các phương pháp khác nhau để đạt hiệu quả tốt nhất.
- Ví dụ:
  - Hàm *np.linalg.svd* trong *NumPy* thực hiện Singular Value Decomposition (SVD).
  - Hàm *sklearn.decomposition.PCA* trong *Scikit-learn* thực hiện Phân tích thành phần chính (PCA).
  - Hàm *sklearn.manifold.TSNE* trong *Scikit-learn* thực hiện T-SNE.
  - Hàm *tensorflow.keras.layers.Embedding* trong *TensorFlow* thực hiện dimensionality reduction cho dữ liệu văn bản.

5. Cung cấp thông tin



Triển khai ứng dụng:

- **Dashboards**: Gồm các biểu đồ, đồ thị và các hình ảnh trực quan dữ liệu. Tương tác với dữ liệu (*lọc dữ liệu theo thời gian, khu vực hoặc sản phẩm*)
- **Reports**: Gồm các tính năng Trực quan hóa dữ liệu, Lọc dữ liệu, Sắp xếp dữ liệu, Xuất dữ liệu.
- **Portals**: Bao gồm dashboards, báo cáo, phân tích dữ liệu và kho dữ liệu.

5. Cung cấp thông tin

6. Tổng kết

Dữ liệu

- **Tâm điểm**: Dữ liệu là tâm điểm của BI. BI thu thập dữ liệu từ nhiều nguồn khác nhau, sau đó xử lý, phân tích và chuyển đổi dữ liệu thành thông tin hữu ích cho doanh nghiệp.
- **Đa dạng**: BI có thể xử lý nhiều loại dữ liệu khác nhau, bao gồm dữ liệu cấu trúc, dữ liệu bán cấu trúc, dữ liệu phi cấu trúc và dữ liệu thời gian thực.
- **Chất lượng**: Chất lượng dữ liệu là yếu tố quan trọng để đảm bảo tính chính xác và hiệu quả của BI.

Question & Answer