

Tổng quan công việc của một Data Scientist - Quy trình triển khai dự án DS-AI

Presenter: Nguyễn Thái Hà - AIVN B2 Team

Supporter: Nguyễn Nhật Linh - AIVN B2 Team

Mục tiêu buổi seminar

- ✓ Dễ dàng hình dung được công việc của một người làm Data Scientist trong môi trường đi làm thực tiễn
- ✓ Hiểu được quy trình triển khai dự án Data Science và AI trong doanh nghiệp từ giai đoạn lập kế hoạch đến giai đoạn vận hành và cải thiện mô hình. Sự khác biệt so với phát triển phần mềm thông thường.
- ✓ Rút ra những bài học thực tế và kinh nghiệm trong việc triển khai dự án Data Science và AI.

Agenda

- I. Giới thiệu bản thân
 - Giới thiệu team AIVN B2
- II. Tổng quan công việc của Data Scientist
 - Data Scientist là ai?
 - Vai trò và trách nhiệm của Data Scientist
 - Các kỹ năng mà Data Scientist cần có
 - Công việc hàng ngày của một data scientist
- III. Quy trình triển khai dự án Data Science và AI
 - Tổng quan và chi tiết quy trình triển khai dự án DS-AI
 - Khác biệt giữa dự án DS-AI và phát triển phần mềm thông thường
 - Các vị trí và cấu trúc team điển hình
 - Giai đoạn 1: Giai đoạn lập kế hoạch
 - Giai đoạn 2: Proof of Concept (PoC)
 - Giai đoạn 3: Triển khai mô hình trên hệ thống
 - Giai đoạn 4: Vận hành, giám sát và cải thiện mô hình
 - Bài học thực tế và kinh nghiệm
- IV. Tổng kết

I. Giới thiệu bản thân

Giới thiệu bản thân



Nguyễn Thái Hà

Quê quán

Hà Nội, Việt Nam

Education

- Đại học Bách Khoa Hà Nội
- Tokyo Institute of Technology, Nhật Bản

Công việc

Data Scientist

A Company

Dự án: Xây dựng hệ thống matching công việc của ứng viên (PoC)

- Lĩnh vực: Tuyển dụng

Dự án: Xây dựng mô hình phân tích độ ăn mòn, dị tật của tank dầu trong nhà máy lọc dầu (PoC)

- Lĩnh vực: Hoá dầu

B Company

Dự án: Xây dựng hệ thống tính toán kho hàng và mô hình dự báo nhu cầu hàng (Hệ thống triển khai trên 15,000 cửa hàng tiện lợi)

- Lĩnh vực: Bán lẻ

Others

- Xây dựng cộng đồng AIVN B2
- Kênh dạy học online (Chia sẻ kiến thức, kinh nghiệm và làm những dự án thực tế)

Giới thiệu AIVN Build Beta (B2) Team

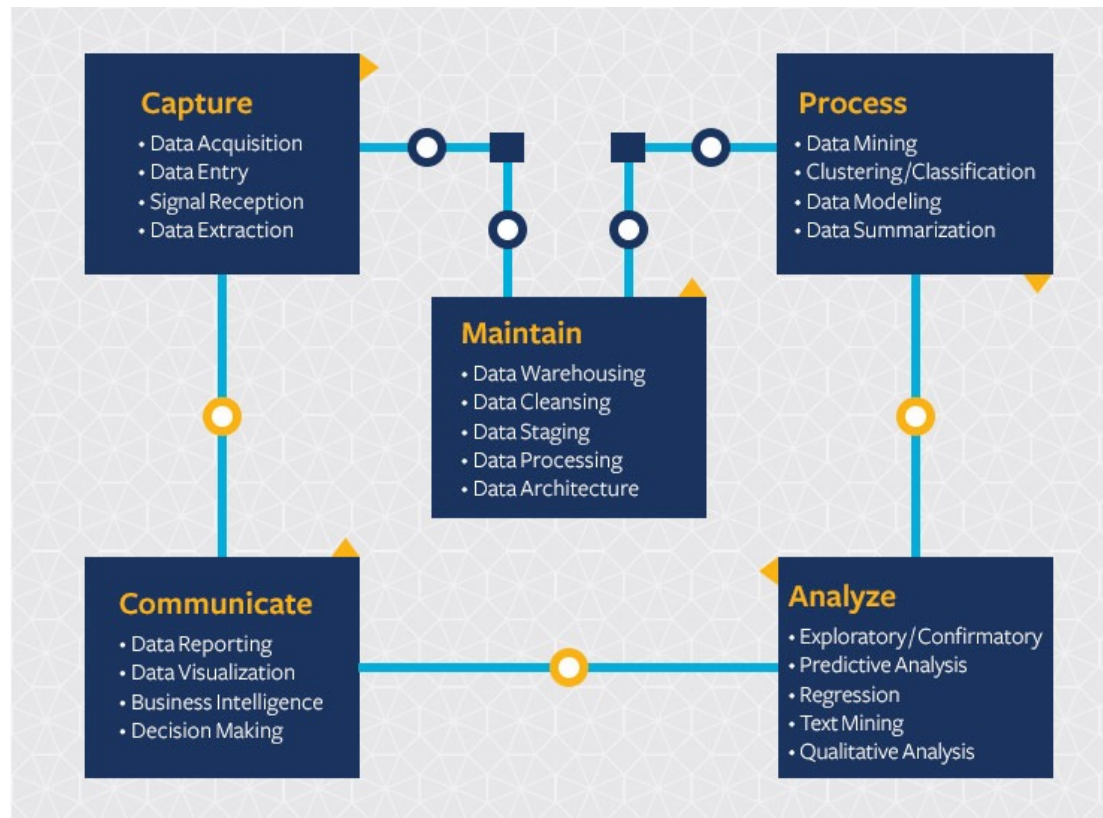


<https://www.youtube.com/watch?v=Ul3pkfK5xlc>

II. Tổng quan công việc của Data Scientist

Data Scientist là ai?

Data Scientist là người có kỹ năng phân tích và lập trình, có khả năng xử lý và tổ chức lượng lớn dữ liệu để tìm ra thông tin hữu ích, hỗ trợ đưa ra quyết định chiến lược cho tổ chức.



Vai trò và trách nhiệm của Data Scientist

Data Scientist chịu trách nhiệm xử lý, phân tích dữ liệu và xây dựng mô hình dự đoán để hỗ trợ các quyết định kinh doanh.

Vai trò chính

Thế mạnh

Data scientist



- Phân tích dữ liệu chuyên sâu để xây dựng mô hình và hỗ trợ đưa ra các quyết định kinh doanh

- Tư duy về dữ liệu tốt
- Khả năng làm việc với nhiều loại dữ liệu
- Kỹ năng xây dựng và tối ưu mô hình AI/ML phức tạp

Data Analyst



- Phân tích dữ liệu để tạo ra các báo cáo
- Cung cấp các insight về dữ liệu để hỗ trợ quyết định kinh doanh

- Khả năng sử dụng các công cụ phân tích như Excel, SQL, và Tableau
- Gần với các dự án kinh doanh
- Khả năng truyền đạt rõ ràng

Data Engineer



- Xây dựng và duy trì hệ thống dữ liệu, ETL pipelines
- Tối ưu hoá và quản lý các kiến trúc dữ liệu Data Base, Data Lake, etc.

- Kỹ năng lập trình mạnh mẽ, am hiểu hệ thống cơ sở dữ liệu, xử lý dữ liệu lớn
- Khả năng xây dựng kiến trúc dữ liệu và đảm bảo dữ liệu được ổn định

ML Engineer

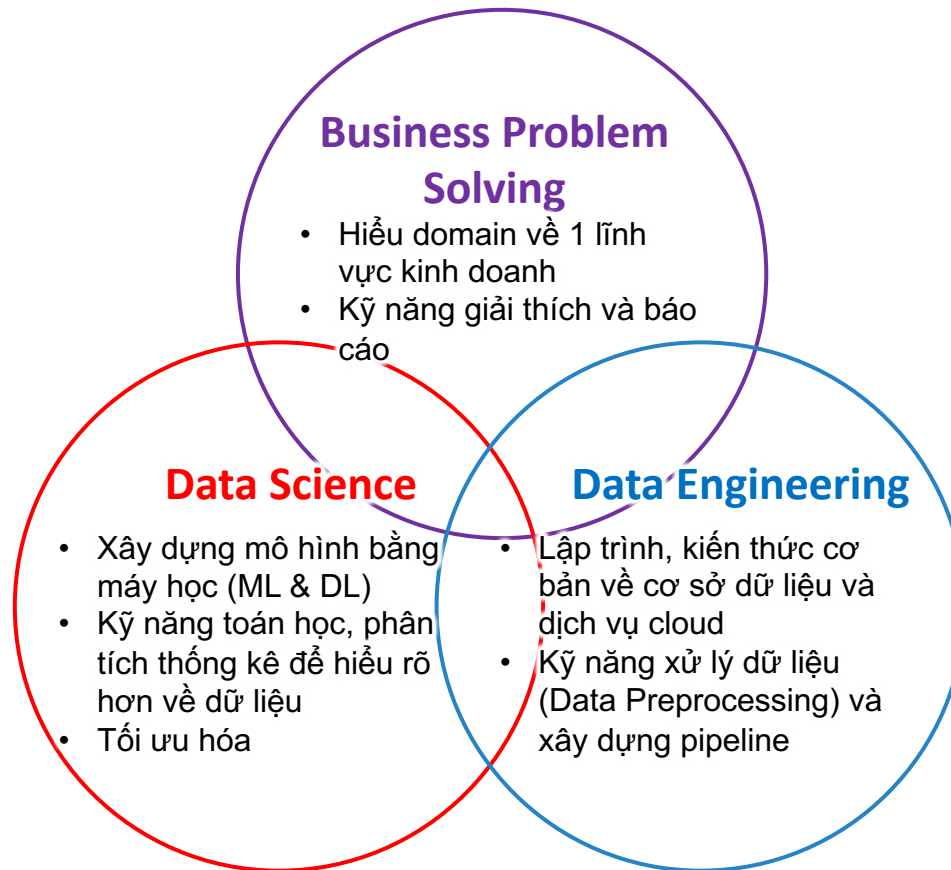


- Triển khai tối ưu hoá và bảo trì các mô hình ML
- Chuyển đổi mô hình từ giai đoạn nghiên cứu sang giai đoạn vận hành

- Kỹ năng lập trình, hiểu biết sâu về các thuật toán ML
- Khả năng triển khai và tích hợp mô hình vào hệ thống thực tế

Các kỹ năng mà Data Scientist cần có

Data Scientist cần có hội tụ ba nhóm kỹ năng chính: giải quyết vấn đề kinh doanh, kỹ thuật xử lý dữ liệu, và kỹ năng về khoa học dữ liệu.



Công việc hàng ngày của một Data Scientist

Công việc DS phụ thuộc vào yêu cầu từng công ty và thường xuyên cần sự phối hợp với các bộ phận khác như IT, Quản lý Dự án, và Khách hàng để đảm bảo các giải pháp AI được triển khai thành công và mang lại giá trị thực sự cho doanh nghiệp.



Thu thập và xử lý dữ liệu

- Tìm kiếm, thu thập và làm sạch dữ liệu từ các nguồn khác nhau.
- Sử dụng các công cụ như SQL, Python, và Spark để quản lý và xử lý dữ liệu.



Phân tích dữ liệu

- Thực hiện phân tích khám phá dữ liệu (EDA) để hiểu các đặc điểm chính.
- Sử dụng các kỹ thuật thống kê và machine learning để phân tích dữ liệu.



Xây dựng mô hình

- Phát triển các mô hình ML để dự đoán, phân loại hoặc phân cụm.
- Tối ưu hóa và điều chỉnh các tham số của mô hình.



Triển khai và giám sát mô hình

- Triển khai mô hình vào môi trường sản xuất.
- Theo dõi hiệu suất của mô hình và cập nhật khi cần thiết.



Tư vấn và cộng tác với khách hàng

- Gặp gỡ và làm việc với khách hàng để hiểu yêu cầu và kỳ vọng.
- Tư vấn về giải pháp AI phù hợp và trình bày kết quả phân tích.



Báo cáo và truyền đạt kết quả

- Chuẩn bị các báo cáo chi tiết và trực quan hóa kết quả.
- Trình bày các phát hiện và đề xuất giải pháp cho nhóm quản lý và khách hàng.

Công việc hàng ngày của một Data Scientist - Case study

Lịch trình làm việc 1 ngày của mình:

☐ Tập trung vào coding phân tích dữ liệu và xây dựng mô hình:



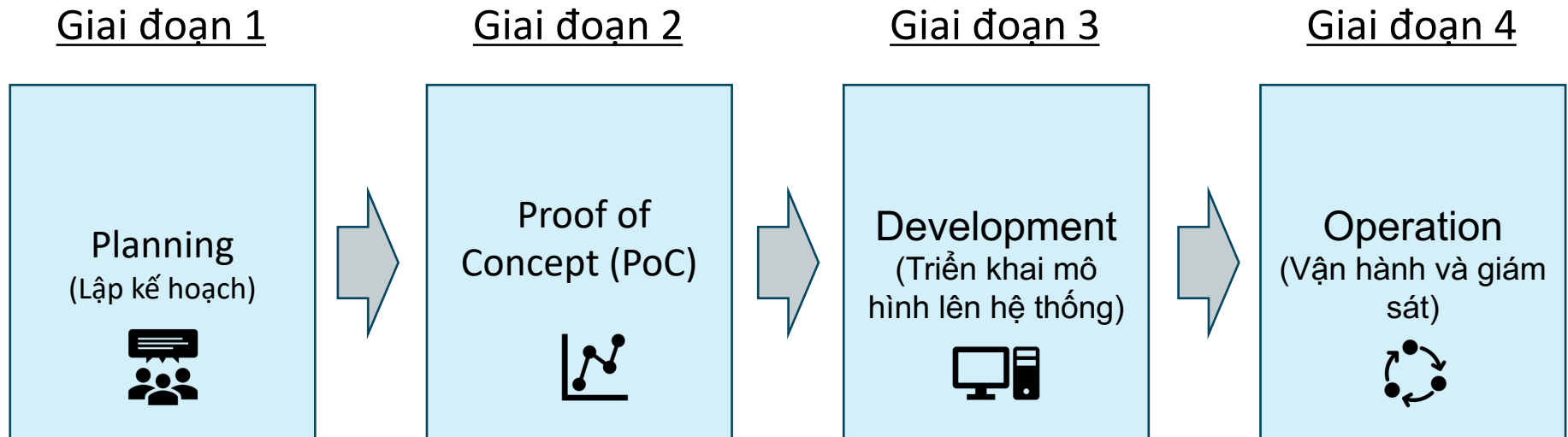
☐ Chuẩn bị báo cáo và báo cáo với khách hàng:



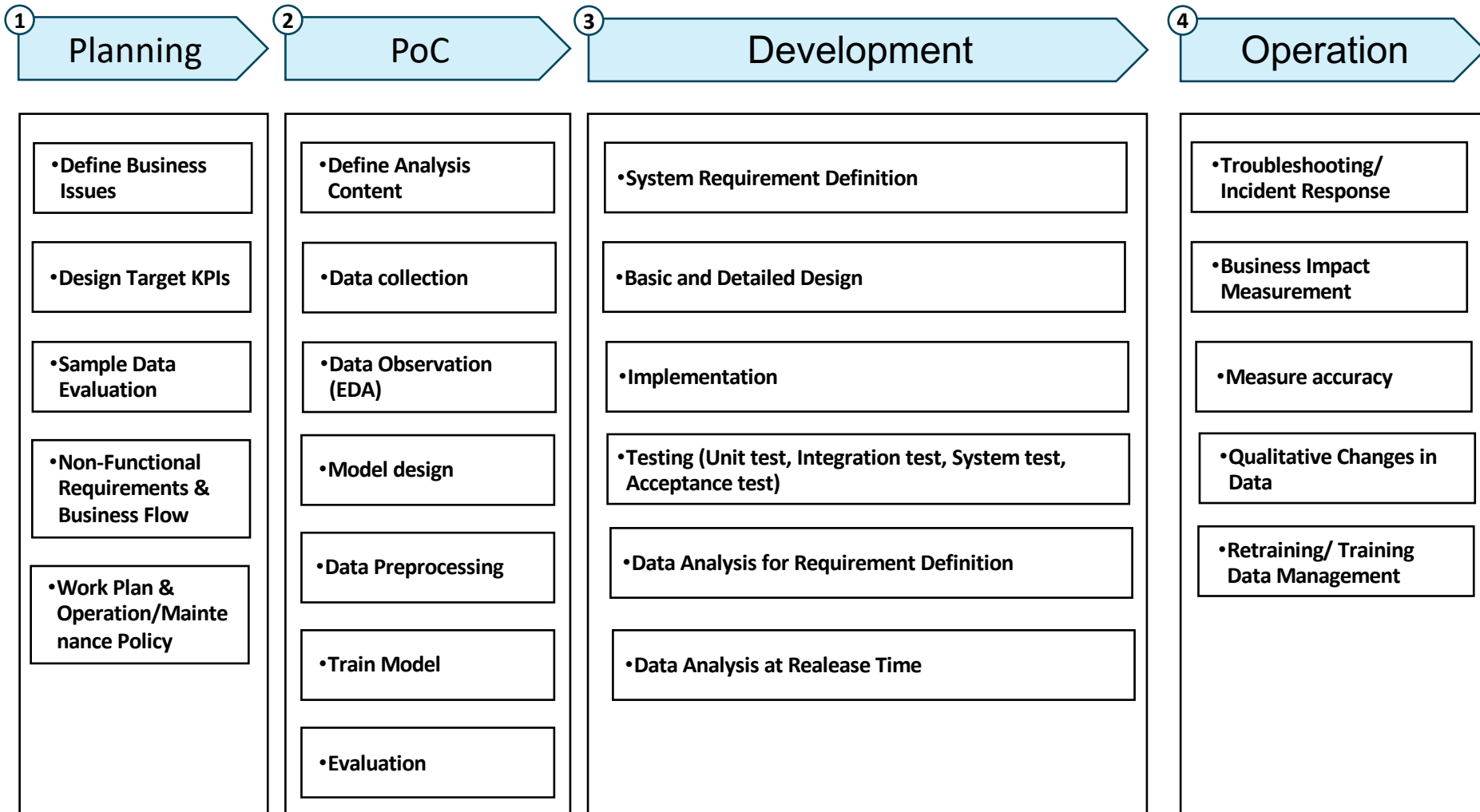
III. Qui trình triển khai dự án Data Science - AI

Tổng quan

Quy trình triển khai dự án DS-AI thường bao gồm 4 giai đoạn: Planning, Proof of Concept (PoC), Development, và Operation.

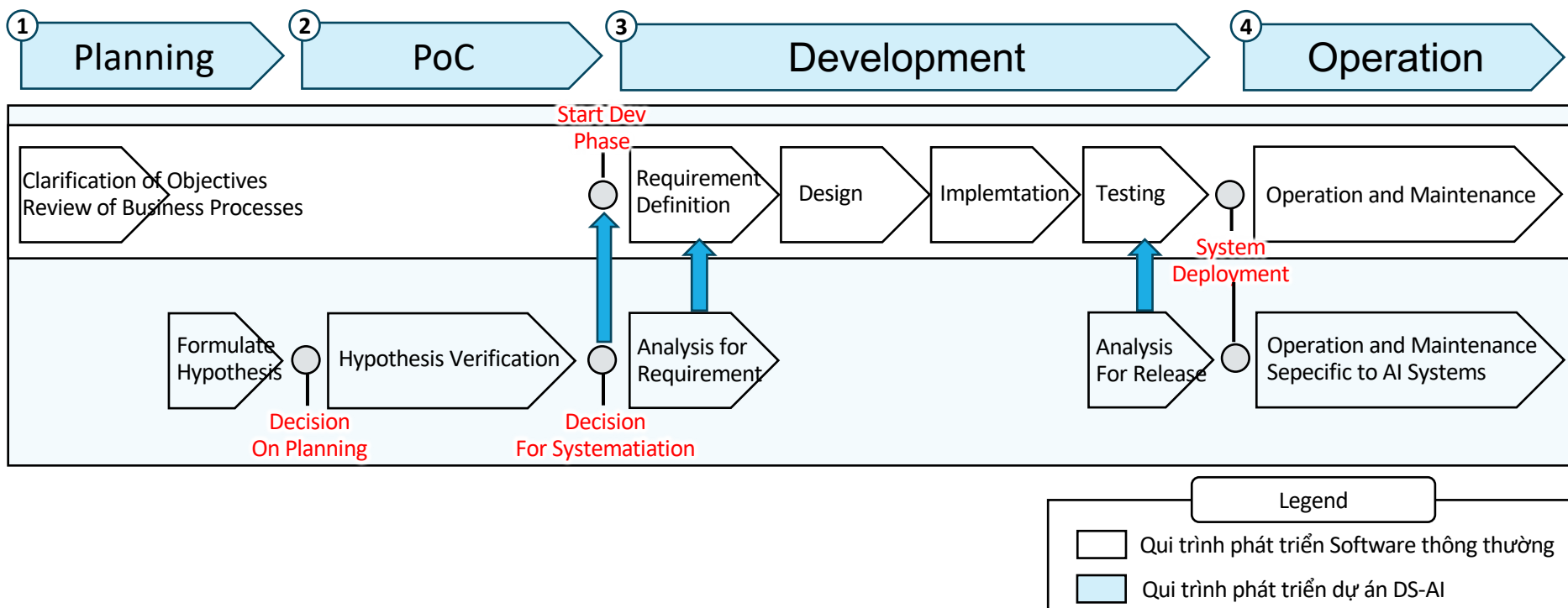


Chi tiết qui trình



Khác biệt giữa dự án DS-AI và phát triển phần mềm thông thường

Quy trình phát triển hệ thống thông thường và quy trình phát triển hệ thống DS-AI có ba điểm khác biệt lớn: có thêm giai đoạn PoC, phân tích dữ liệu, và quy trình vận hành đặc thù.

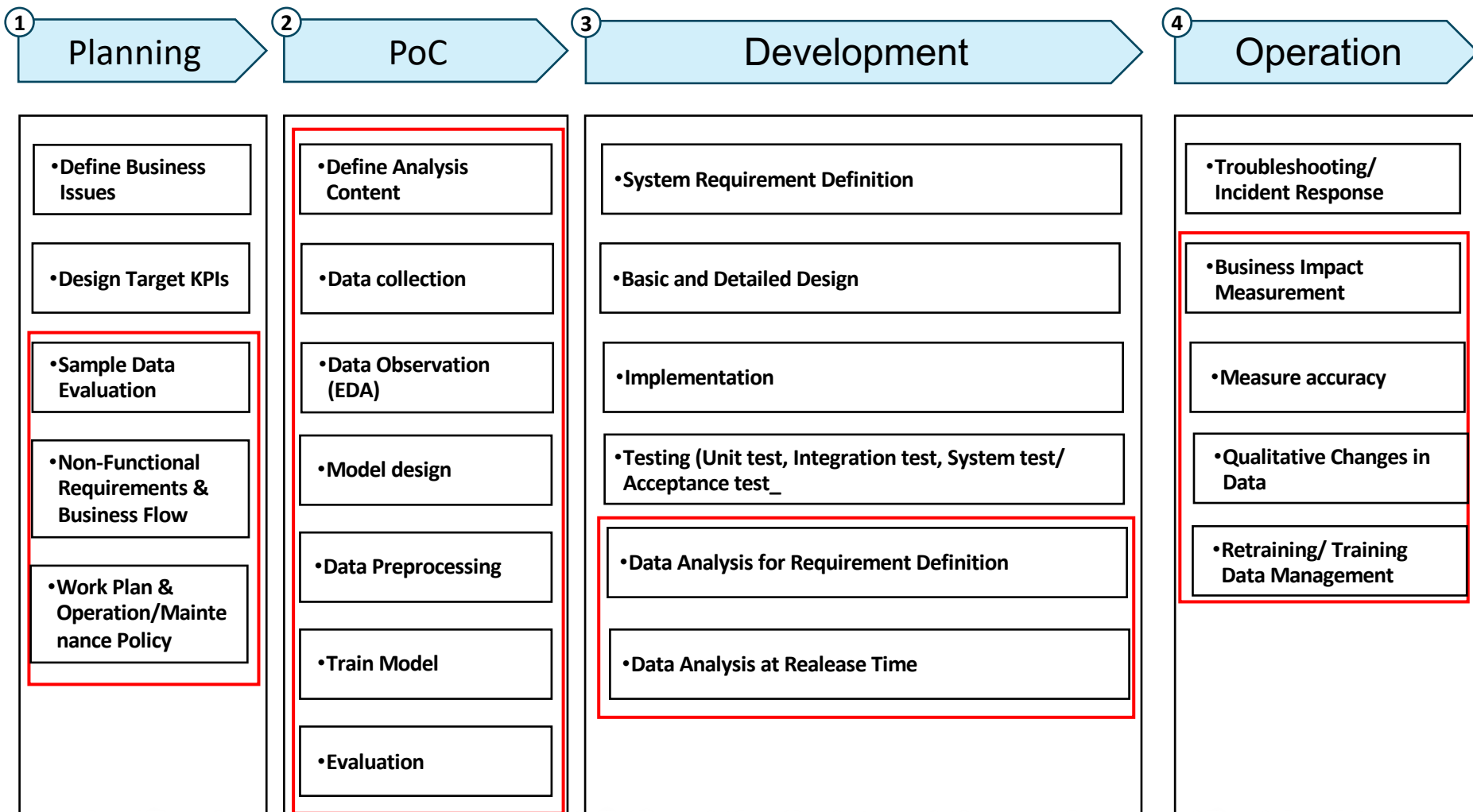


Các vị trí cần thiết trong 1 dự án DS-AI

Một dự án cần có 2 vị trí chính: member & leader. Các nhiệm vụ cụ thể rõ ràng đảm bảo dự án được triển khai và vận hành hiệu quả.

<u>Member</u>		<u>Leader</u>	
Title	Nhiệm vụ chính	Title	Nhiệm vụ chính
Data Analyst/ Data Scientist	<ul style="list-style-type: none">• Phân tích dữ liệu, tiền xử lý dữ liệu• Xây dựng mô hình, đánh giá mô hình• Đưa ra các yêu cầu phân tích	Data Analyst Team Lead	<ul style="list-style-type: none">• Quản lý nhóm phân tích dữ liệu• Xác định, giám sát các KPIs• Định hướng chiến lược phân tích dữ liệu
Machine Learning Engineer	<ul style="list-style-type: none">• Thiết kế, xây dựng tối ưu mô hình• Triển khai mô hình, retraining• Xử lý, chuẩn bị dữ liệu training	Developer Team Lead	<ul style="list-style-type: none">• Quản lý và dẫn dắt nhóm Dev• Đảm bảo chất lượng phần mềm và hiệu suất hệ thống và quy trình test
Developer	<ul style="list-style-type: none">• Lập trình, triển khai và tích hợp các thành phần của hệ thống• Thực hiện các quy trình testing	Consultant	<ul style="list-style-type: none">• Cung cấp các tư vấn chiến lược• Hỗ trợ trong việc xác định yêu cầu và lập kế hoạch cho dự án
Researcher	<ul style="list-style-type: none">• Nghiên cứu, thử nghiệm các công nghệ mới liên quan đến DS-AI• Cập nhật xu hướng áp dụng vào dự án	Researcher Team Lead	<ul style="list-style-type: none">• Lãnh đạo nhóm nghiên cứu, định hướng nghiên cứu và thử nghiệm công nghệ mới
		Project Manager (PM)	<ul style="list-style-type: none">• Quản lý toàn bộ dự án, đảm bảo tiến độ và chất lượng dự án• Điều phối giữa các nhóm khác nhau từ phân tích dữ liệu, phát triển đến nghiên cứu• Đảm bảo dự án đáp ứng yêu cầu của khách hàng và KPIs

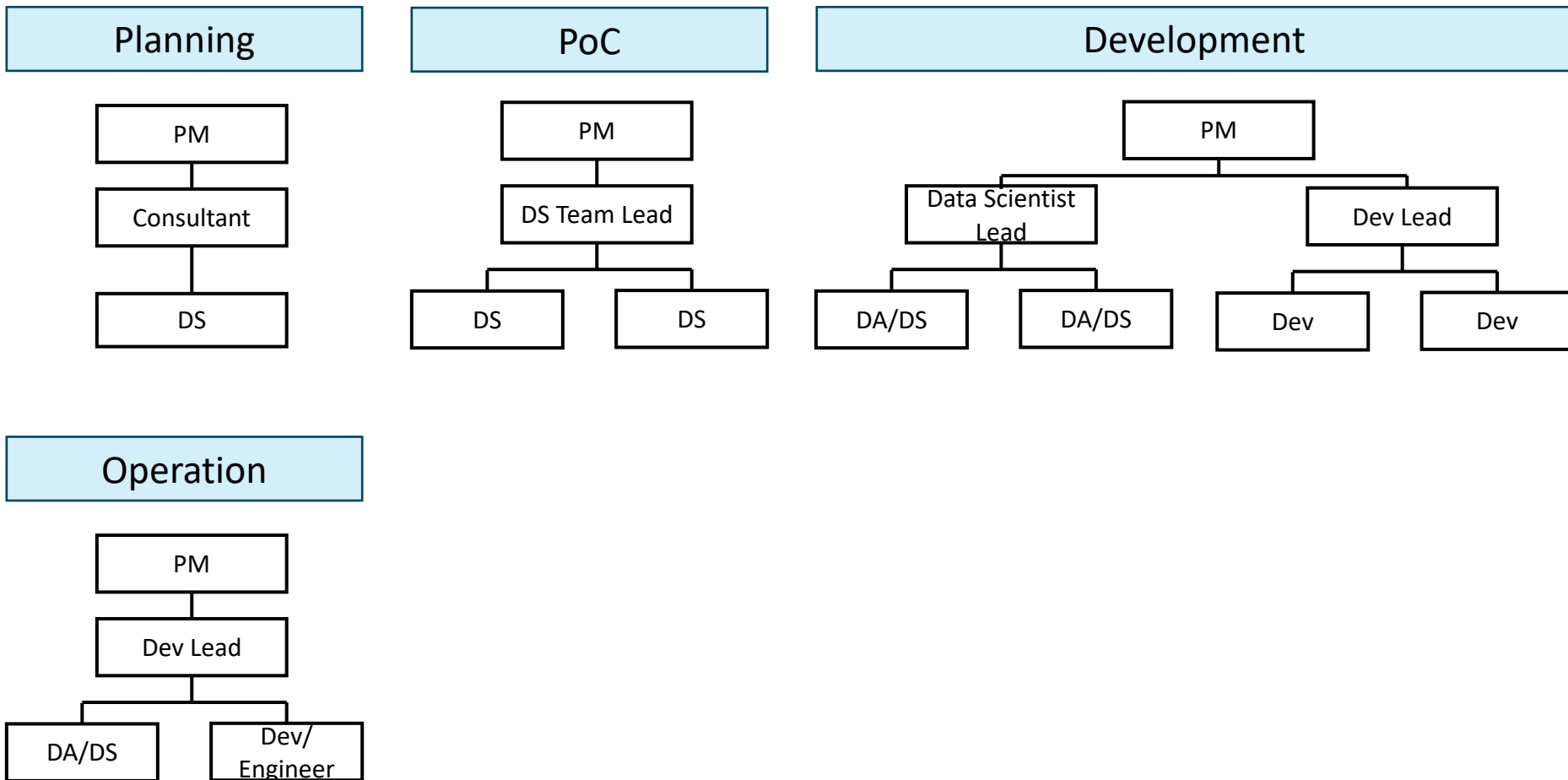
Công việc Data Scientist có thể đảm nhiệm trong dự án DS-AI



Các phần công việc Data Scientist có thể đảm nhiệm là những nhiệm vụ liên quan đến phân tích dữ liệu, đánh giá mẫu dữ liệu, tiền xử lý dữ liệu, thiết kế mô hình, và đánh giá tác động kinh doanh trong các giai đoạn chính của dự án DS-AI

Ví dụ cấu trúc team theo từng giai đoạn

Sơ đồ tổ chức và phân chia vai trò trong nhóm phát triển dự án DS-AI

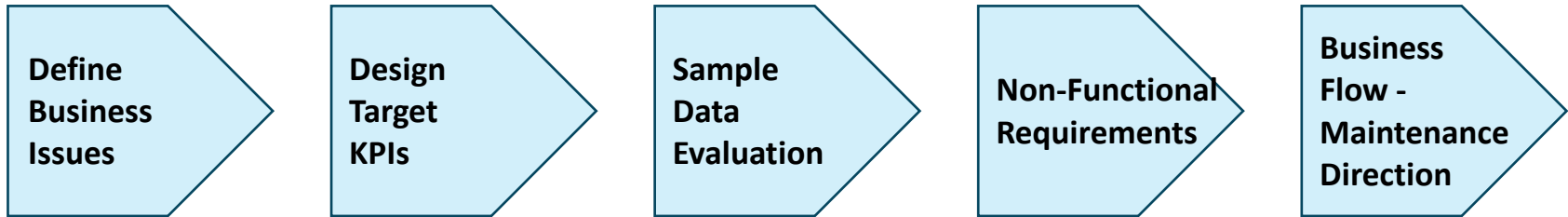


Cách phân chia team sẽ phụ thuộc rất nhiều vào công ty và quy mô cũng như độ khó của dự án.

Giai đoạn 1: Lập kế hoạch

Chi tiết qui trình

Thông thường giai đoạn lên kế hoạch thường sẽ kéo dài từ 1-2 tháng và bao gồm 5 bước chính.



Mục tiêu	Xác định và hiểu rõ vấn đề kinh doanh cần giải quyết <ul style="list-style-type: none">✓ Mục tiêu này giúp đảm bảo rằng dự án được thiết kế để sử dụng AI giải quyết đúng vấn đề cần thiết, từ đó tạo ra giá trị cho doanh nghiệpDùng thử AI, hoặc đưa AI vào hệ thống không phải là mục tiêu tốt
	Thiết kế các chỉ số KPI mục tiêu để đo lường thành công <ul style="list-style-type: none">✓ Đặt ra các KPI rõ ràng và có thể đo lường để đánh giá hiệu quả của giải pháp, giúp hướng dẫn và theo dõi quá trình phát triển dự ánXây dựng mô hình LightGBM với độ chính xác >90% không phải là KPI tốt
	Đánh giá khả năng và chất lượng của dữ liệu có sẵn <ul style="list-style-type: none">✓ Mục tiêu này tập trung vào việc xác định dữ liệu hiện tại có đủ chất lượng và sẵn sàng để thực hiện các phân tích cần thiết, từ đó đưa ra quyết định chính xác về các bước tiếp theo trong dự ánMột số công ty có sẵn dữ liệu nội bộ tuy nhiên dữ liệu có sẵn này có thể không phù hợp với yêu cầu của mô hình

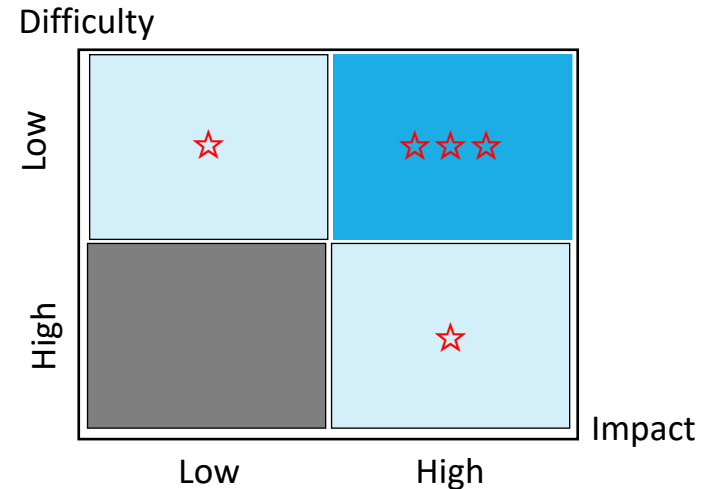
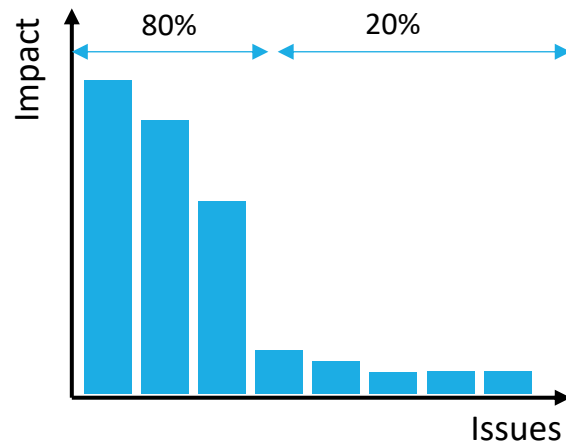
Define Business Issues - Làm sao để xác định đúng vấn đề

Để có một dự án Data Science mang lại hiệu quả cao nhất thì việc xác định rõ các vấn đề từ đó đưa ra phương hướng giải quyết và thiết kế đúng KPIs rất quan trọng.



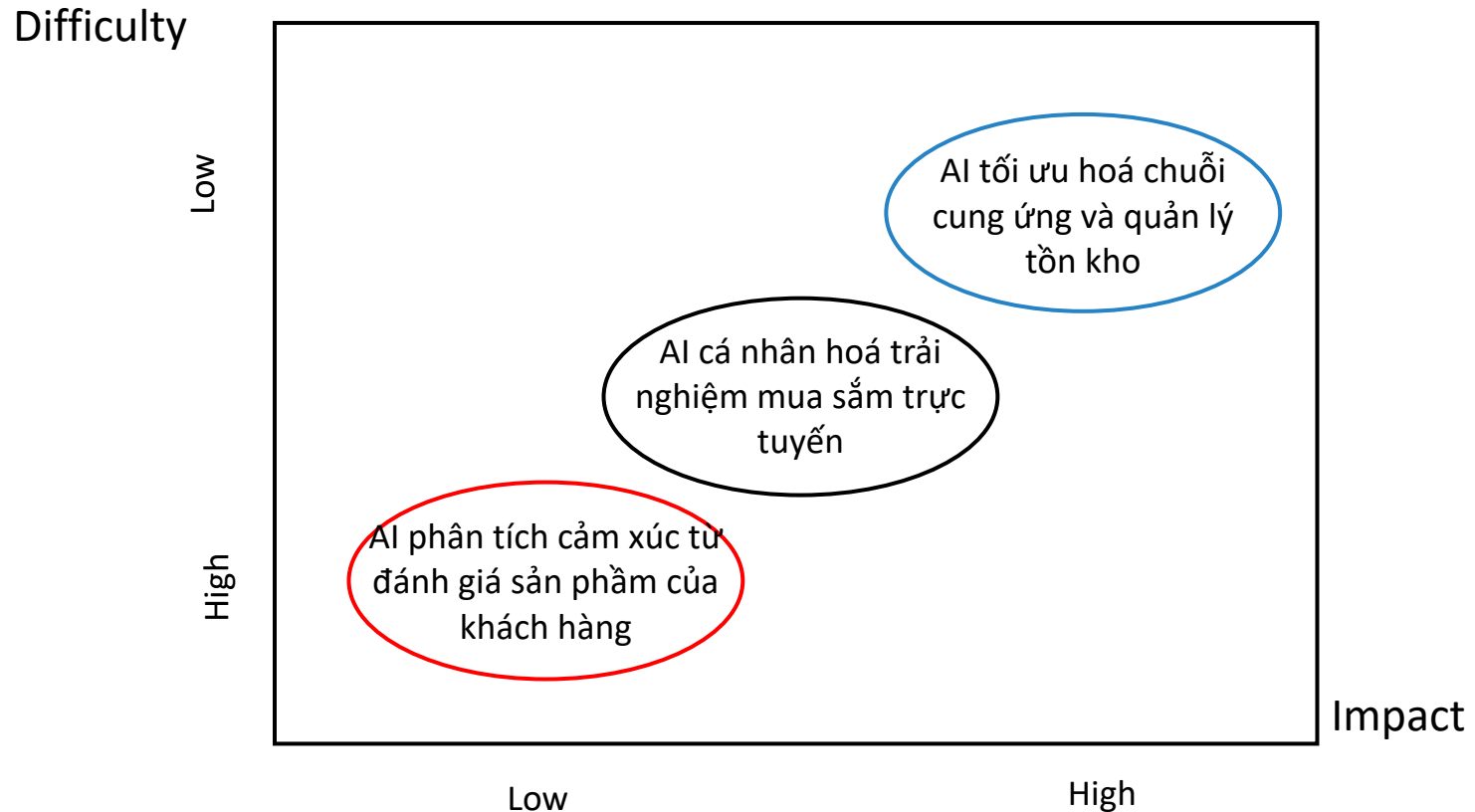
Phương pháp xác định vấn đề rõ ràng

- ✓ Chinh lý thông tin, lắng nghe nhiều ý kiến từ những phòng ban liên quan
- ✓ Liệt kê và đánh giá tất cả các vấn đề có thể tồn tại
- ✓ Áp dụng phương pháp 80:20 hoặc Order of Magnitude để định lượng và đánh giá độ ảnh hưởng thứ tự ưu tiên của các vấn đề



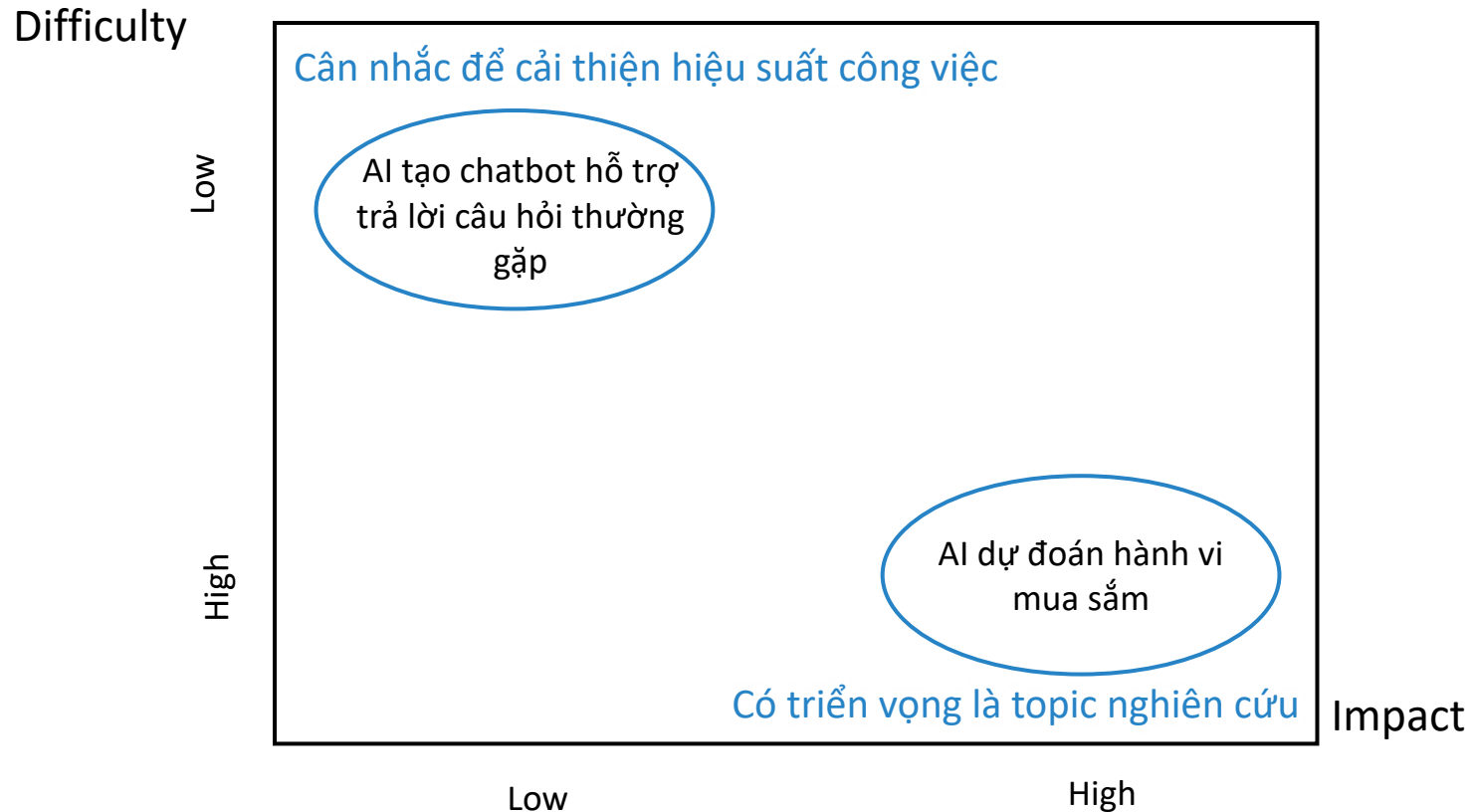
Làm sao để xác định đúng vấn đề - Case study 1

Xem xét hai yếu tố quan trọng: Tác động (Impact) và Độ khó (Difficulty) để lựa chọn dự án khả thi và có giá trị nhất.



Làm sao để xác định đúng vấn đề - Case study 2

Xem xét hai yếu tố quan trọng: Tác động (Impact) và Độ khó (Difficulty) để lựa chọn dự án khả thi và có giá trị nhất.



Design Target KPIs

Việc thiết kế mục tiêu và KPIs rõ ràng giúp doanh nghiệp đo lường được hiệu quả của hệ thống AI. Đảm bảo đạt được mục tiêu kinh doanh đã đề ra.


Case study

Kế hoạch	Xây dựng AI tối ưu hoá chuỗi cung ứng và quản lý tồn kho
Mục tiêu	Tối ưu hoá chuỗi cung ứng và quản lý tồn kho để giảm thiểu lãng phí và tối đa hoá hiệu quả vận hành
KPIs	Giảm tỷ lệ tồn kho dư thừa xuống dưới 10% trong vòng 6 tháng sau khi triển khai hệ thống AI
	Giảm 15% chi phí lưu kho hàng tháng
	Đạt tỷ lệ chính xác dự báo nhu cầu hàng hoá trên 85%
	Giảm chi phí vận chuyển và giao hàng liên quan đến chuỗi cung ứng xuống 10%

Sample Data Evaluation

Quy trình đánh giá dữ liệu giúp đảm bảo dữ liệu đủ tốt và phù hợp cho việc phát triển mô hình AI.

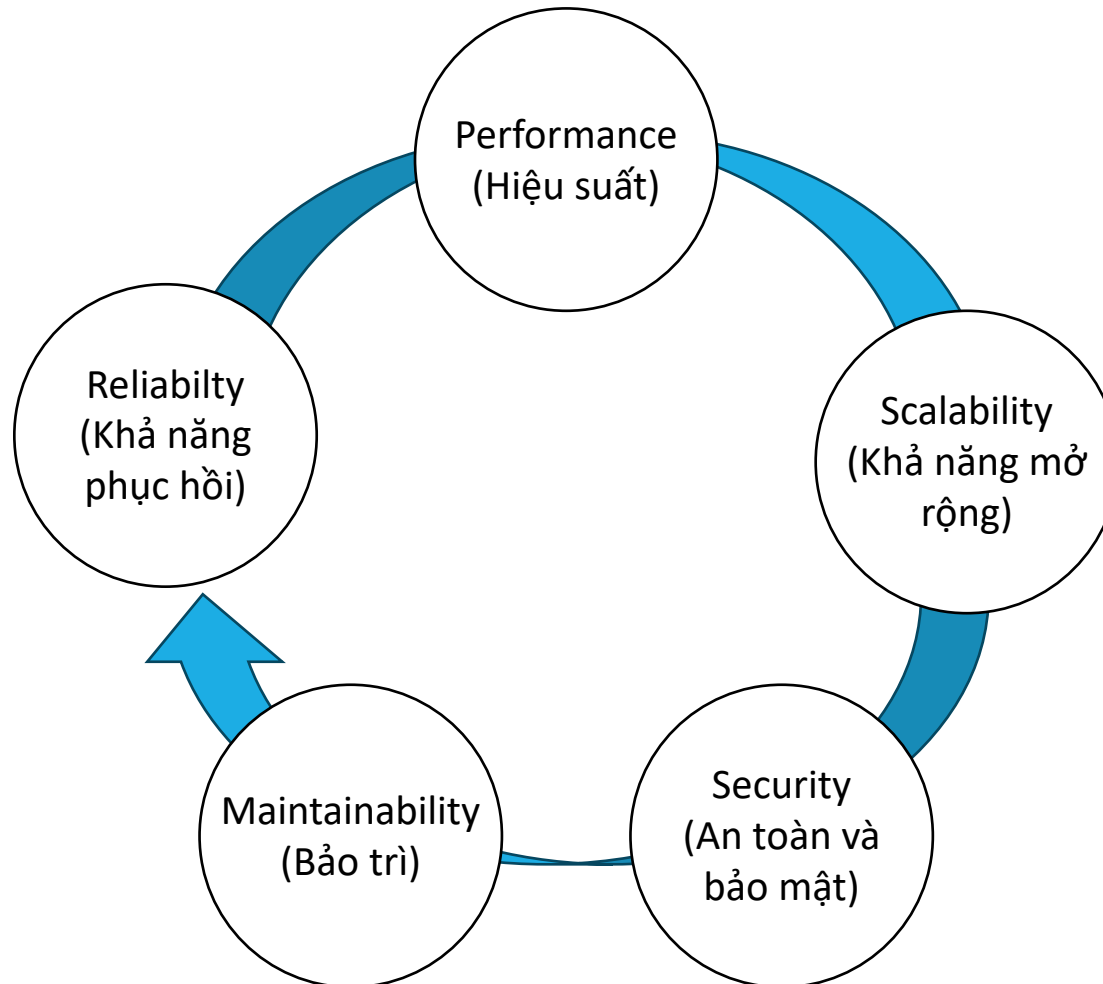
Mục tiêu	<ul style="list-style-type: none">Xác định chất lượng và tính khả dụng của dữ liệu để đảm bảo rằng dữ liệu đủ tốt để xây dựng mô hình AI.
	<ul style="list-style-type: none">Đánh giá các vấn đề tiềm ẩn như thiếu dữ liệu, dữ liệu không chính xác, hoặc dữ liệu không đồng nhất.



Kết quả mong đợi	<ul style="list-style-type: none">Báo cáo đánh giá chất lượng dữ liệu, liệt kê các vấn đề tìm thấy và đề xuất phương pháp xử lý.
	<ul style="list-style-type: none">Quyết định về việc tiếp tục sử dụng dữ liệu hiện có hoặc yêu cầu thêm dữ liệu.

Non-Functional Requirements

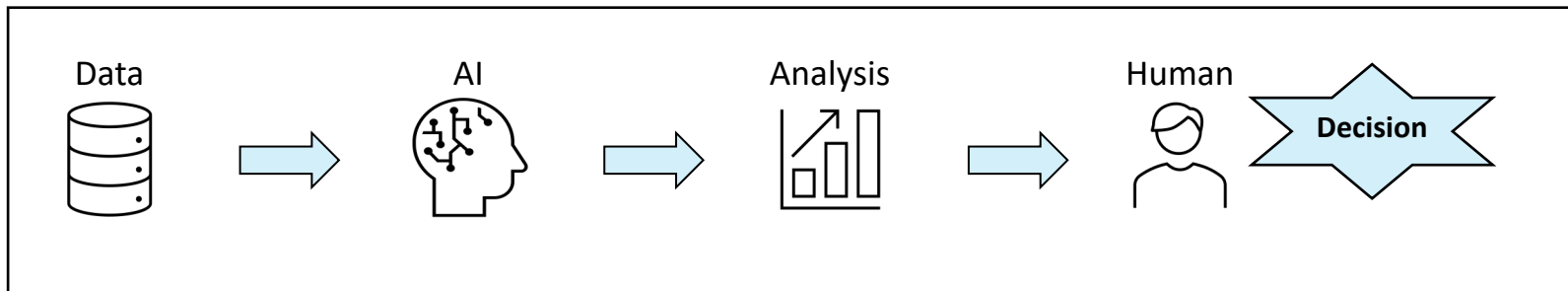
Định nghĩa các yêu cầu phi chức năng cho hệ thống AI nhằm đảm bảo hệ thống hoạt động ổn định, an toàn và đáp ứng các kỳ vọng về hiệu suất.



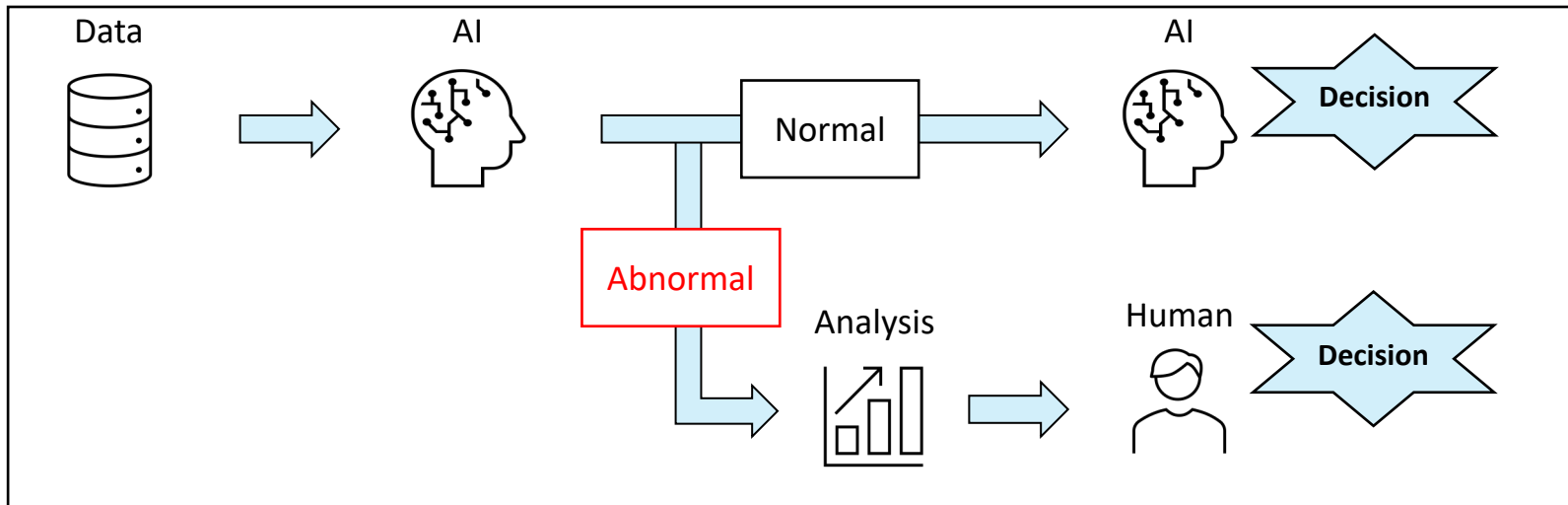
Business Flow

Trong giai đoạn planning, cần thiết kế rõ ràng work flow và định nghĩa vai trò của con người và AI.

☐ Con người nắm quyền đưa ra quyết định cuối cùng



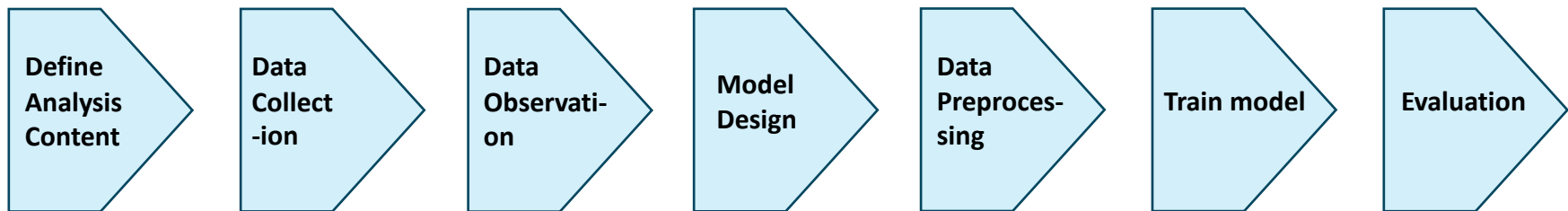
☐ Kết hợp linh hoạt giữa AI và con người để tự động or bán tự động



Giai đoạn 2: Proof of Concept (PoC)

Tổng quan và chi tiết quy trình

PoC là quá trình thử nghiệm việc đưa dữ liệu vào mô hình AI để kiểm tra xem có đạt được mục tiêu như dự kiến hay không. Giai đoạn này thường kéo dài từ 2-3 tháng tùy vào độ khó của dự án.



Mục tiêu	• Xác định rõ giá trị dự kiến mà hệ thống AI cần đạt được
	• Kiểm chứng giả thuyết mô hình AI có thể giải quyết được vấn đề đã đề ra
	• Đánh giá khả năng của mô hình trong việc đáp ứng các yêu cầu thực tế trước khi triển khai hệ thống chính thức

Define Analysis Content (Định nghĩa nội dung phân tích)

Ở giai đoạn này cần chọn đối tượng phân tích chính cho PoC để đánh giá khả năng AI có thể đạt được mục tiêu dự án mà không phải phân tích toàn bộ dữ liệu.

Case study

Mục tiêu	Sử dụng AI để dự báo doanh thu và tối ưu quy trình đặt hàng cho chuỗi cửa hàng tiện lợi trên toàn quốc <ul style="list-style-type: none">Dự báo lượng nhu cầu các sản phẩm với độ chính xác caoTự động quyết định lượng hàng tối ưu dựa trên kết quả dự báo chính xác
Đối tượng phân tích	10 cửa hàng (Chọn 10 cửa hàng có doanh thu cao và ổn định) 50 sản phẩm (chọn các sản phẩm thông thường có tỷ lệ doanh thu lớn)
Dữ liệu	1 năm gần nhất
Thời gian dự báo	Doanh thu và đặt hàng cho 2 tuần tới (Từ ngày N+1 -> N+14)
Đơn vị dự báo	Dự báo hàng giờ (1 giờ)

Data Collection (Thu thập và tổng hợp dữ liệu)

Nhiệm vụ chính là xác định và thu thập dữ liệu cần thiết cho mô hình dự đoán trong giai đoạn PoC



Xác định nguồn dữ liệu

- Dữ liệu có thể bao gồm dữ liệu lịch sử, dữ liệu thời gian thực, hoặc dữ liệu từ bên thứ ba. Đảm bảo rằng nguồn dữ liệu được chọn có thể cung cấp thông tin phù hợp với mục tiêu của dự án.



Định nghĩa các biến mục tiêu (target variable) và biến giải thích (features)

- Biến mục tiêu là số liệu mà bạn muốn dự đoán hoặc tối ưu hóa.
- Biến giải thích là các yếu tố có khả năng ảnh hưởng đến biến mục tiêu và cần được thu thập để đưa vào mô hình dự đoán.



Xác nhận tính khả dụng của dữ liệu

- Kiểm tra và xác nhận rằng dữ liệu có thể được thu thập và sử dụng trong quá trình vận hành thực tế của hệ thống. Điều này giúp đảm bảo rằng hệ thống có thể hoạt động ổn định khi chuyển từ môi trường thử nghiệm sang môi trường thực tế.



Xử lý dữ liệu thiếu (NA values) hoặc bất thường (Outliers)

- Xác định và xử lý các vấn đề liên quan đến dữ liệu thiếu hoặc bất thường. Đây là bước quan trọng để đảm bảo chất lượng dữ liệu đầu vào cho mô hình.

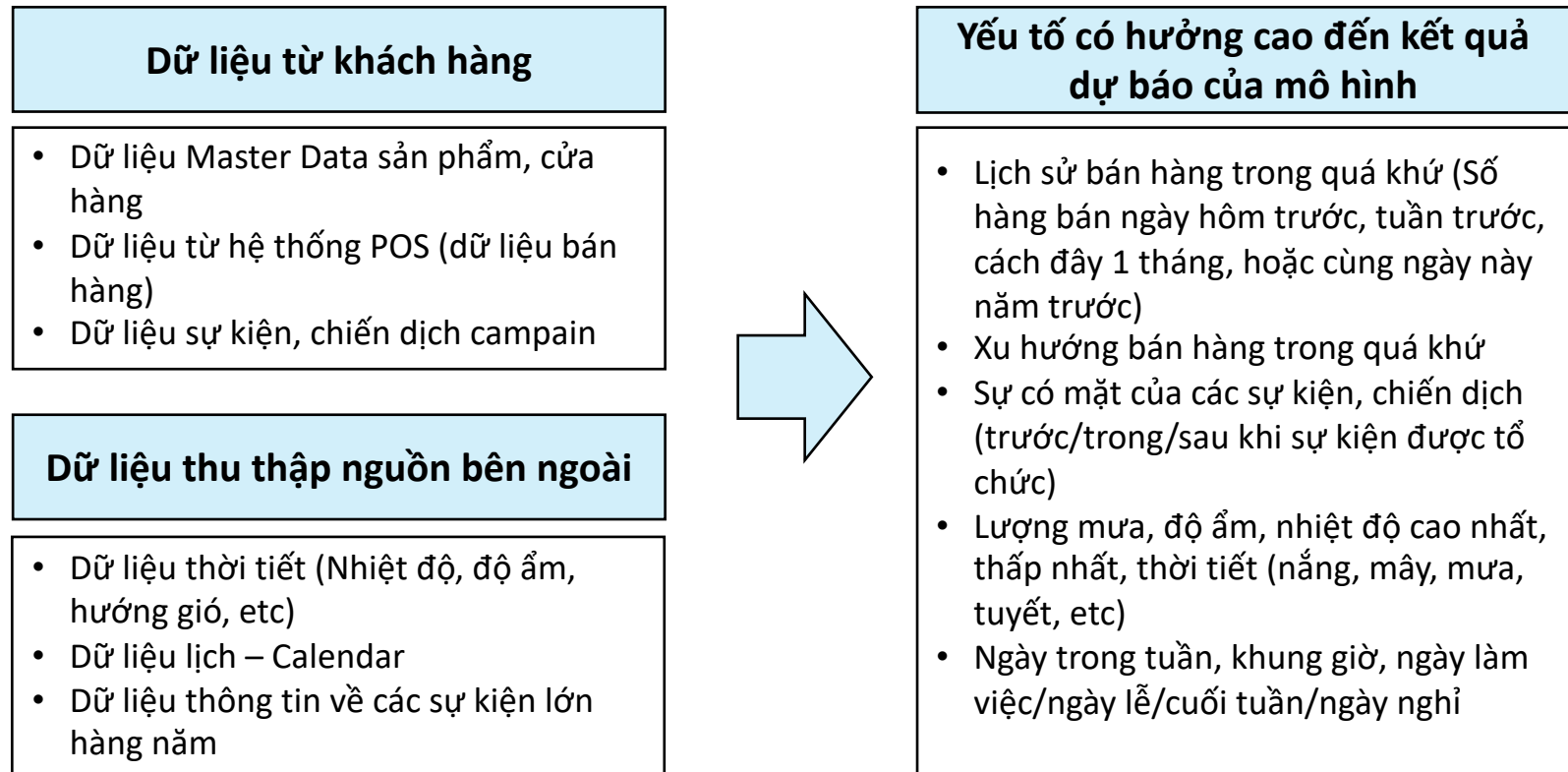


Tích hợp và lưu trữ dữ liệu

- Tích hợp dữ liệu từ nhiều nguồn khác nhau và lưu trữ chúng một cách an toàn và có tổ chức để sử dụng trong các bước tiếp theo của dự án.

Data Collection (Thu thập và tổng hợp dữ liệu) – Case study

Dưới đây là một số loại dữ liệu có thể sử dụng cho bài toán dự báo doanh thu của chuỗi cửa hàng tiện lợi



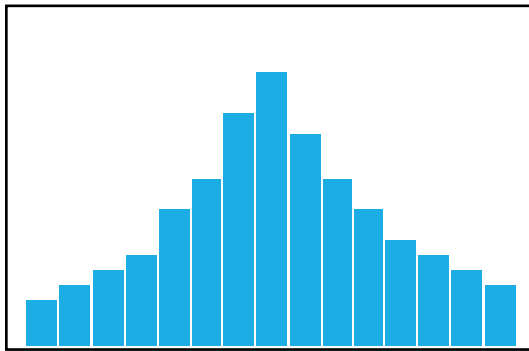
Data Observation (Quan sát dữ liệu) (1/2)

Thông thường dữ liệu raw data sẽ phải được xử lý trước khi đưa vào mô hình AI. Do đó việc kiểm tra dữ liệu trước khi xử lý là rất quan trọng. Khi quan sát dữ liệu, có thể phát hiện một số vấn đề từ đó đưa ra các biện pháp cải thiện kết quả.

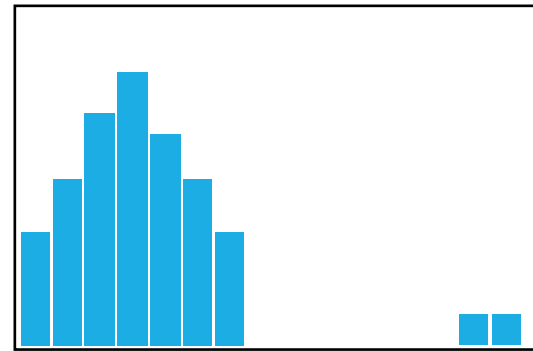
Loại dữ liệu	Hạng mục cần kiểm tra
Dữ liệu số Numerical Data	✓ Kiểm tra phân bố của độ phân tán và trung bình (Biểu đồ histogram)
	✓ Giá trị bất thường (outliers)
	✓ Tỷ lệ dữ liệu thiếu
Categorical data	✓ Kiểm tra sự cần thiết của việc phân loại lại các danh mục
	✓ Kiểm tra sự đồng đều trong việc xử lý các dữ liệu khác ngoài thang đo danh nghĩa
	✓ Giá trị bất thường và tỷ lệ dữ liệu thiếu
Dữ liệu ảnh	✓ Màu sắc, kích thước có đồng đều hay không
	✓ Kích thước và vị trí của các đối tượng cần chú ý trong ảnh có nhất quán không
	✓ Xem xét sự tồn tại của các yếu tố gây nhiễu trong ảnh như noise, artifact
Dữ liệu text	✓ Kiểm tra các lỗi chính tả, ngữ pháp hoặc ký tự đặc biệt không mong muốn
	✓ Với văn bản tự do hoặc không có cấu trúc, độ dài câu có đồng đều không
	✓ Tính nhất quán của ngữ cảnh trong các văn bản để đảm bảo không có sự mâu thuẫn trong thông tin

Data Observation (Quan sát dữ liệu) (2/2)

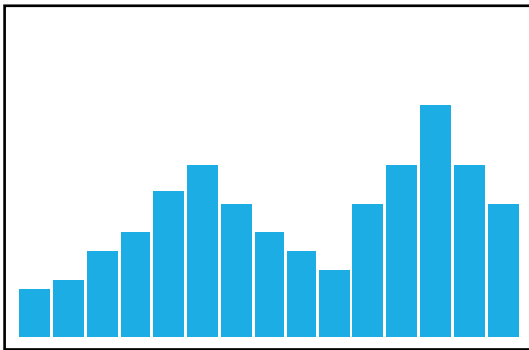
Sử dụng biểu đồ histogram để quan sát phân bố của dữ liệu số một cách trực quan. Kết hợp thêm việc tính toán các giá trị thống kê đặc trưng như giá trị trung bình, trung vị, độ phân tán sẽ hữu ích.



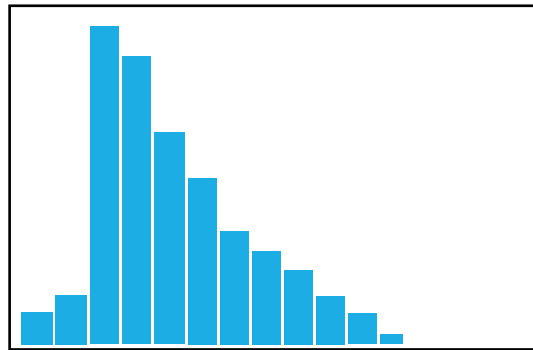
Phân phối chuẩn



Phân phối chuẩn (có outliers)



Phân phối chuẩn hỗn hợp



Phân bố lệch phải

Model Design (Thiết kế mô hình)

Khi thiết kế mô hình việc lựa chọn thuật toán là bước quan trọng. Dựa trên yêu cầu của bài toán và đặc điểm của dữ liệu, các loại thuật toán khác nhau sẽ được sử dụng.

Loại thuật toán	Hồi quy	Phân loại	Trường hợp thích hợp	Trường hợp không thích hợp
Deep learning	<input type="radio"/>	<input type="radio"/>	<ul style="list-style-type: none"> Cần độ chính xác cao Bộ dữ liệu đủ lớn Xử lý dữ liệu hình ảnh hoặc text 	<ul style="list-style-type: none"> Cần giải thích và hiểu rõ nguyên nhân của kết quả dự đoán Bộ dữ liệu nhỏ
XGBoost Light GBM	<input type="radio"/>	<input type="radio"/>	<ul style="list-style-type: none"> Cần độ chính xác cao Dữ liệu dạng bảng 	<ul style="list-style-type: none"> Cần giải thích và hiểu rõ nguyên nhân của kết quả dự đoán
Random Forest	<input type="radio"/>	<input type="radio"/>	<ul style="list-style-type: none"> Cần độ chính xác tương đối cao Cần giải thích kết quả dự đoán 	<ul style="list-style-type: none"> Khi muốn nhóm đối tượng theo các thuộc tính khác nhau
Decision Tree	<input type="radio"/>	<input type="radio"/>	<ul style="list-style-type: none"> Muốn giải thích lý do mô hình đưa ra dự đoán 1 cách đơn giản Muốn nhóm đối tượng theo thuộc tính 	<ul style="list-style-type: none"> Yêu cầu độ chính xác cao Có nhiều yếu tố ảnh hưởng phức tạp
Logistic Regression	<input type="radio"/>	<input type="radio"/>	<ul style="list-style-type: none"> Khi muốn giải thích lý do phân loại và các biến quan trọng cùng lúc 	<ul style="list-style-type: none"> Cần độ chính xác cao Dữ liệu phức tạp
Linear Regression	<input type="radio"/>	<input type="radio"/>	<ul style="list-style-type: none"> Khi giá trị của biến mục tiêu quan trọng và cần giải thích kết quả dự đoán 	<ul style="list-style-type: none"> Cần độ chính xác Dữ liệu phức tạp



QUIZ TIME

Theo hướng của cột “Loại thuật toán” từ trên xuống và từ dưới lên có quy luật như thế nào? 35

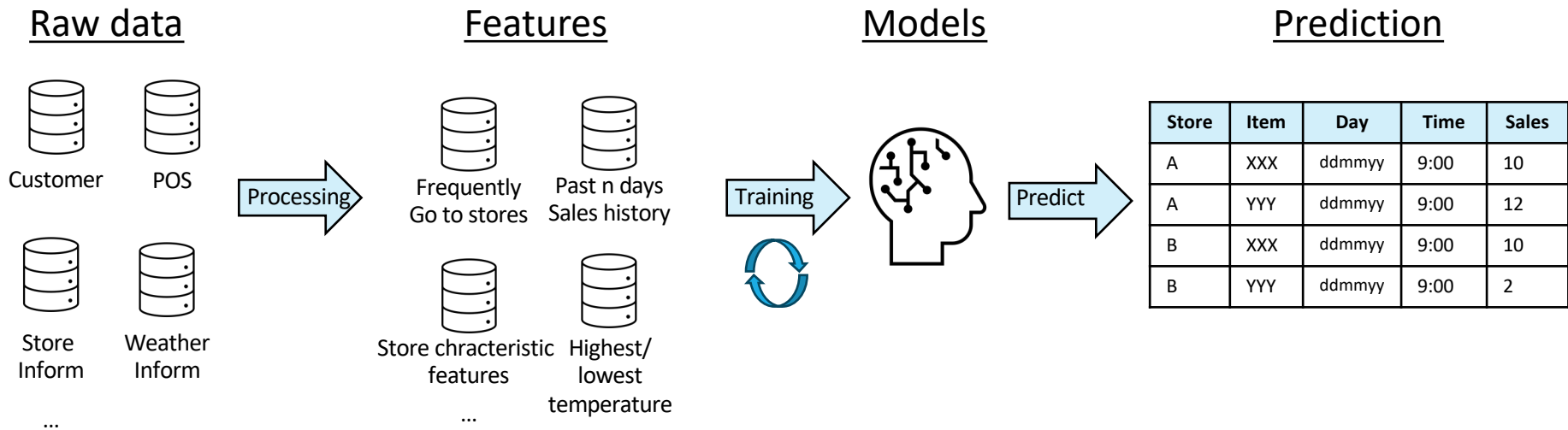
Data Preprocessing (Xử lý dữ liệu)

Việc xử lý và làm sạch dữ liệu trước khi sử dụng dữ liệu để xây dựng mô hình chiếm phần lớn thời gian và khối lượng công việc. Tuy nhiên đây là giai đoạn rất quan trọng và đòi hỏi sự tỉ mỉ và kiến thức về lĩnh vực kinh doanh cụ thể (domain knowledge).

Hạng mục xử lý	Nội dung chi tiết
Làm sạch dữ liệu	✓ Đánh giá chất lượng dữ liệu: giá trị thiếu, trùng lặp, không nhất quán, bất thường
	✓ Xử lý dữ liệu thiếu bằng cách bổ sung, loại bỏ dữ liệu trùng lặp, chỉnh sửa lại hạng mục bị sai
	✓ Kiểm tra tính nhất quán: Có mẫu thuẫn trong dữ liệu không (Về thời gian, or logic)
Biến đổi dữ liệu	✓ Tạo feature đặc trưng (Feature Engineering): Tạo ra các đặc trưng mới từ dữ liệu hiện có
	✓ Tổng hợp dữ liệu(groupby-aggregation): Tổng hợp dữ liệu dựa trên yêu cầu kinh doanh (Eg.daily sales)
	✓ Data Transformation: biến đổi log, normalization, scaling, label encoding, OHE.
Tích hợp nhiều nguồn dữ liệu	✓ Kết hợp các tập dữ liệu: Kết hợp nhiều tập dữ liệu từ các nguồn khác nhau
	✓ Kết hợp nhiều loại dữ liệu: dữ liệu từ nhiều định dạng khác nhau (văn bản, ảnh, dữ liệu cảm biến)
Xác thực dữ liệu	✓ Data Integrity Checks: Xác thực tính toàn vẹn của dữ liệu sau khi biến đổi và tích hợp
	✓ Xác thực format, schema: xác định rằng dữ liệu tuân thủ theo format, schema mong đợi
	✓ Kiểm tra mẫu dữ liệu: lấy mẫu ngẫu nhiên or có tiêu chuẩn để kiểm tra chất lượng dữ liệu, so sánh các mẫu dữ liệu trước và sau khi xử lý để xác thực không có thông tin quan trọng bị mất

Train model (Huấn luyện mô hình)

Sau khi xử lý làm sạch dữ liệu và tạo những features mới, chúng ta đã sẵn sàng cho việc xây dựng mô hình. Xây dựng mô hình là quá trình try and fail. Từ Feedback của mô hình có thể giúp điều chỉnh lại bộ features.



Evaluation (Đánh giá mô hình)

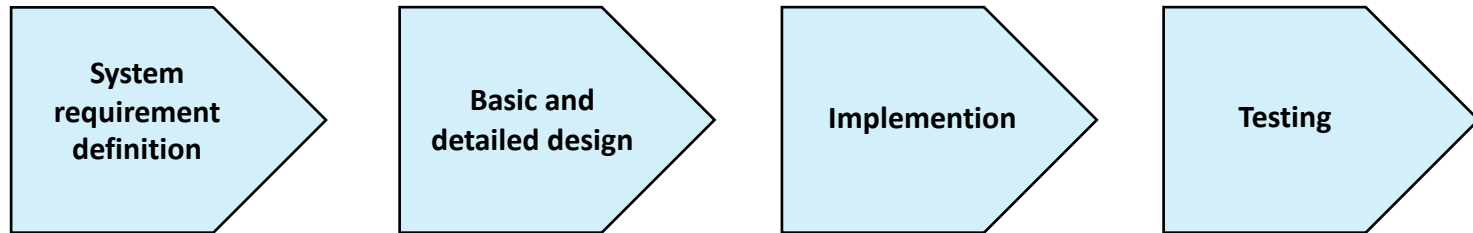
Các tiêu chí đánh giá mô hình thường bao gồm độ chính xác, khả năng giải thích, mức độ overfitting, thời gian thực thi.

Độ chính xác	<ul style="list-style-type: none">Là chỉ đo lường kết quả học của mô hình dựa trên tập dữ liệu validation và test.
Khả năng giải thích	<ul style="list-style-type: none">Đánh giá dựa trên việc kết quả dự đoán của mô hình có dễ hiểu hay không, liệu con người có thể lý giải và đủ tin tưởng sử dụng kết quả này không
Mức độ overfitting	<ul style="list-style-type: none">Đánh giá xem kết quả dự đoán của mô hình có bị overfitting không (Cho độ chính xác rất cao trên tập training, tuy nhiên không có performance tốt trên tập validation or test)
Thời gian thực thi	<ul style="list-style-type: none">Đo lường thời gian cần thiết để thực hiện quá trình học, vì điều này có thể là một vấn đề khi đưa vào hệ thống

Giai đoạn 3: Development

Quy trình và mục tiêu của giai đoạn phát triển hệ thống

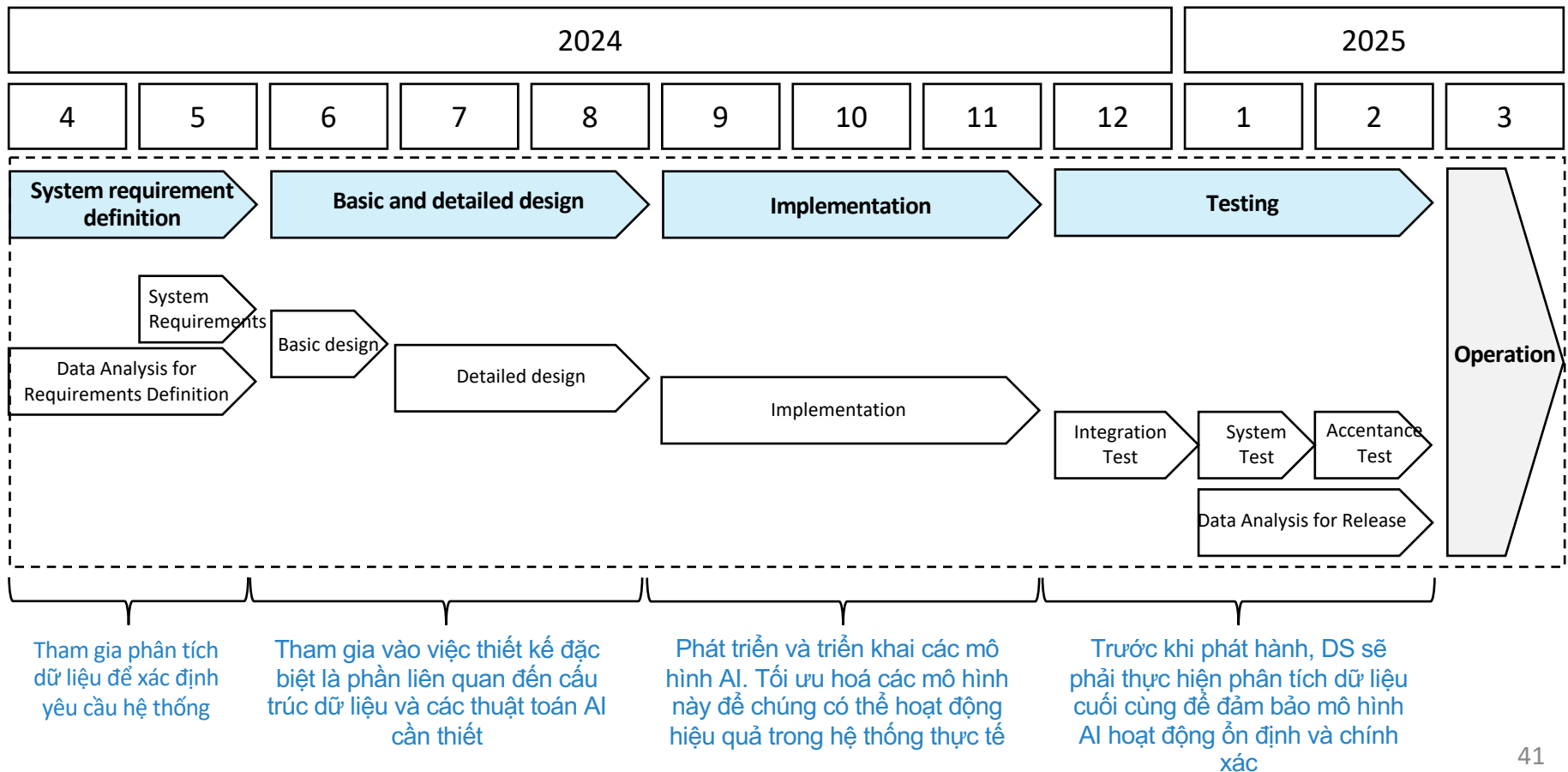
Mục tiêu của giai đoạn phát triển là xây dựng, tối ưu hoá hiệu suất và kiểm thử hệ thống AI hoàn chỉnh.



Mục tiêu	Thiết kế và xây dựng hệ thống AI hoàn chỉnh <ul style="list-style-type: none">✓ Phát triển một hệ thống AI hoàn chỉnh từ việc thu thập và xử lý dữ liệu, đến việc triển khai mô hình AI đã được kiểm chứng trong giai đoạn PoC.✓ Hệ thống phải đáp ứng được các yêu cầu kỹ thuật nghiệp vụ đã đề ra giai đoạn trước
	Đảm bảo hiệu suất và tính ổn định của hệ thống <ul style="list-style-type: none">✓ Tối ưu hoá hiệu suất của hệ thống AI để đảm bảo nó hoạt động hiệu quả, xử lý dữ liệu nhanh chóng và duy trì tính ổn định trong quá trình vận hành
	Kiểm thử và đảm bảo chất lượng <ul style="list-style-type: none">✓ Thực hiện các quy trình kiểm thử để đảm bảo rằng hệ thống AI hoạt động chính xác, không có lỗi và đáp ứng được các tiêu chí chất lượng.

Lịch trình và vai trò DS đảm nhiệm

Giai đoạn phát triển là 1 chuỗi các hoạt động từ **xác định yêu cầu, thiết kế triển khai và kiểm thử** để đảm bảo **hệ thống AI hoạt động hiệu quả ổn định**. Trong giai đoạn này **DS đóng vai trò cầu nối quan trọng giữa lý thuyết và thực tiễn**, đảm bảo rằng mô hình AI không chỉ chính xác mà còn sẵn sàng cho việc triển khai và vận hành.



Giai đoạn 4: Operation

Mục tiêu và quy trình tương ứng

Giai đoạn vận hành và bảo trì tập trung vào việc duy trì hoạt động ổn định của hệ thống, cải tiến liên tục mô hình AI và bảo vệ hệ thống khỏi các rủi ro bảo mật, đảm bảo rằng hệ thống luôn đáp ứng được các yêu cầu nghiệp vụ và kỹ thuật

Mục tiêu	Đảm bảo hệ thống AI hoạt động ổn định và hiệu quả <ul style="list-style-type: none">✓ Giám sát liên tục (Continuous Monitoring): Thực hiện giám sát hệ thống AI trong thời gian thực để phát hiện và xử lý kịp thời các lỗi hoặc sự cố có thể xảy ra.✓ Quản lý tài nguyên (Resource Management): Tối ưu hoá việc sử dụng tài nguyên hệ thống như CPU, GPU, bộ nhớ và dung lượng lưu trữ để hệ thống hoạt động mượt mà và hiệu quả
	Cập nhật và cải tiến mô hình AI <ul style="list-style-type: none">✓ Thu thập dữ liệu thực tế (Real world data collection): Thu thập dữ liệu mới từ các hoạt động của hệ thống để cập nhật và tinh chỉnh mô hình AI, đảm bảo rằng mô hình liên tục học hỏi và cải thiện theo thời gian✓ Huấn luyện lại mô hình (Retraining): Thực hiện huấn luyện lại mô hình AI định kỳ hoặc khi có dữ liệu mới quan trọng
	Quản lý rủi ro và bảo mật hệ thống <ul style="list-style-type: none">✓ Kiểm tra và cập nhật bảo mật (Security Audits and Updates): Kiểm tra bảo mật thường xuyên để phát hiện xác lỗ hổng và triển khai các bản vá bảo mật để bảo vệ hệ thống AI khỏi các mối đe dọa✓ Quản lý phiên bản (Version Control): Duy trì quản lý các phiên bản của mô hình AI và hệ thống

Bài học thực tế và kinh nghiệm



Mục tiêu và KPI rõ ràng

- **Mục tiêu cụ thể:** Xác định rõ ràng mục tiêu của PoC là gì, ví dụ như kiểm chứng tính khả thi của một thuật toán, đo lường hiệu suất của mô hình trên dữ liệu thực tế, hay đánh giá sự tương thích của một giải pháp với hệ thống hiện có.
- **Kết quả mong đợi:** Đặt ra các chỉ số đo lường rõ ràng (KPI) để đánh giá thành công của PoC, ví dụ như độ chính xác của mô hình, thời gian xử lý, hoặc khả năng mở rộng của giải pháp.



Phối hợp chặt chẽ với các bên liên quan (Team working and communication)

- **Phối hợp liên tục với các bên liên quan:** Làm việc chặt chẽ với các đội ngũ khác như kinh doanh, pháp lý, và vận hành từ giai đoạn phát triển đến vận hành để hiểu rõ nhu cầu và đảm bảo tính khả thi của dự án.
- **Hiểu sâu sắc nhu cầu người dùng:** Sự phối hợp này giúp tối ưu hóa mô hình AI để đáp ứng chính xác nhu cầu của người dùng cuối, dẫn đến tăng hiệu suất và thành công của hệ thống khi đưa vào vận hành.



Tham gia vào dự án thực tế sớm

- **Lời khuyên:** Không gì có thể thay thế được kinh nghiệm thực tiễn. Hãy tìm kiếm các cơ hội để tham gia vào các dự án thực tế, dù đó là dự án cá nhân, dự án nhóm với bạn bè, hay tham gia các cuộc thi như Kaggle. Điều này giúp bạn áp dụng những gì đã học vào thực tế và hiểu rõ hơn về quy trình làm việc trong lĩnh vực này.
- **Tại sao quan trọng:** Thực hiện các dự án giúp bạn tích lũy kinh nghiệm, xây dựng portfolio cá nhân và chuẩn bị tốt hơn cho công việc trong tương lai.

IV. Tổng kết

Tổng kết



Công việc của Data Scientist

- ✓ **Vai trò:** Data Scientist đóng vai trò trong việc xử lý và phân tích dữ liệu, phát triển mô hình AI và đảm bảo mô hình hoạt động hiệu quả trong thực tế
- ✓ **Kỹ năng:** Công việc của Data Scientist không chỉ đòi hỏi kiến thức về toán học và khoa học dữ liệu, mà còn cần sự hiểu biết về nghiệp vụ và khả năng làm việc nhóm.



Quy trình triển khai dự án DS-AI

- ✓ **Giai đoạn:** Bao gồm 4 giai đoạn chính từ việc lập kế hoạch, kiểm chứng mô hình PoC, đến phát triển và vận hành.
- ✓ **Khác biệt so với dự án phát triển phần mềm thông thường:** Quy trình triển khai dự án DS-AI có thêm giai đoạn PoC, phân tích dữ liệu, và quy trình vận hành đặc thù.
- ✓ **Vai trò của Data Scientist:** Data Scientist đảm nhiệm vai trò quan trọng trong các giai đoạn chính của dự án DS-AI (Eg. Phân tích dữ liệu, đánh giá mẫu, tiền xử lý dữ liệu, thiết kế mô hình, và đánh giá tác động kinh doanh trong)

Tài liệu tham khảo

Chưa có bản
dịch tiếng Việt

