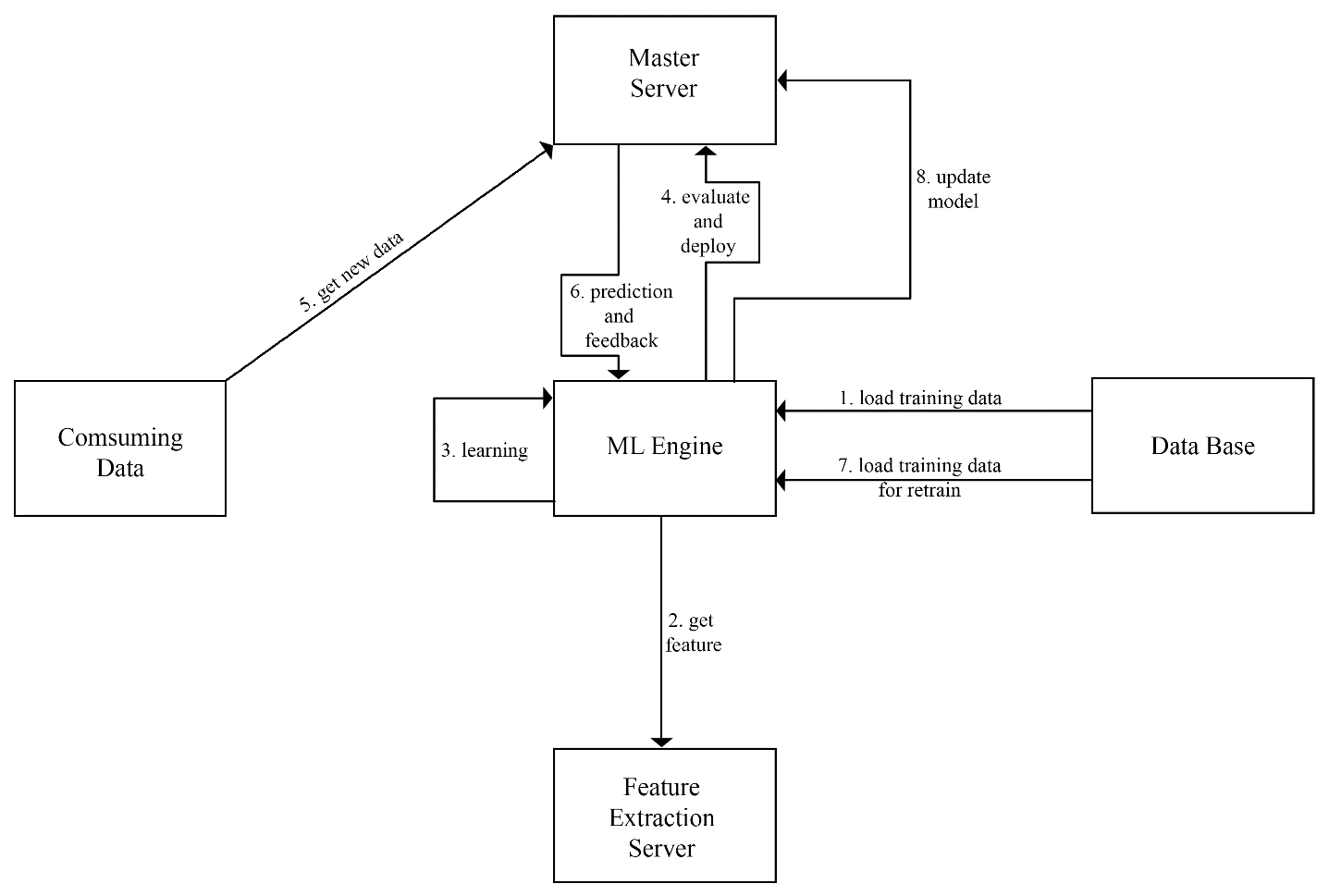
Malware Detection Framework

1. Bài toán

Phát hiện các tệp tin PE độc (malware) trong hệ thống.

1. Mô hình



1) Load dữ liệu huấn luyện từ cơ sở dữ liệu. Dữ liệu huấn luyện là tập các tệp tin đã được đánh nhãn độc (malware) hoặc sạch (benign).

2) Trích chọn đặc trưng cho mỗi tệp tin trong tập huấn luyện – vector hóa các tệp tin, mỗi tệp tin sẽ được biểu diễn bởi vector tương ứng, các vector này được gọi là vector đặc trưng. Các đặc trưng bao gồm: api, n-grams, entropy histogram.

3) Xây dựng mô hình học máy với dữ liệu huấn luyện là các vector đặc trưng được trích chọn ở 2. Dữ liệu được huấn luyện bởi Deep Neural Networks.

4) Đánh giá lựa chọn mô hình và deploy mô hình.

5) Lấy dữ liệu mới từ hệ thống – các tệp tin PE.

6) Dự đoán nhãn cho tệp tin đến, verify kết quả và feed back.

7) Load dữ liệu huấn luyện cũ trong cơ sở dữ liệu, kết hợp với dữ liệu feed back phục vụ cho quá trình huấn luyện lại (retrain).

8) Đánh giá mô hình trên tập dữ liệu feed back, cập nhật mô hình mới nếu độ chính xác tốt hơn mô hình cũ.