Project Design Phase-I

Solution Architecture

Date	20 October 2023
Team ID	2.5
Project Name	Malware Detection and Classification
Team Members	• Hardik Kankane • Anondita Dutta • Elisabeth Ann Varghese

1. Introduction

1.1 Purpose

The purpose of this document is to provide a comprehensive overview of the solution architecture for the Malware Detection and Classification system.

1.2 Scope

This document covers the architectural design, components, data flow, technology stack, deployment architecture, security measures, compliance considerations, and other relevant aspects of the system.

1.3 Audience

This document is intended for technical stakeholders, including architects, developers, system administrators, and security professionals involved in the design, development, deployment, and maintenance of the Malware Detection and Classification system.

1.4 Document Overview

This document is structured into several sections, each addressing specific aspects of the solution architecture. It begins with an introduction, followed by a system overview, functional and non-functional requirements, data flow, technology stack, deployment architecture, monitoring, security measures, compliance considerations, integration points, scalability, and future considerations.

2. System Overview

2.1 High-Level Architecture

The Malware Detection and Classification system is designed as a distributed, cloud-native application. It comprises several components working together to provide comprehensive malware detection and classification capabilities.

2.2 Components Overview

- **Data Ingestion Module**: Responsible for collecting and ingesting data from various sources, such as files, network traffic, and endpoints.
- **Pre-processing Module**: Cleans and prepares the data for further analysis, including data normalization and feature extraction.
- **Machine Learning Model**: Utilizes a pre-trained model for malware detection and classification. This module is responsible for making predictions based on the input data.
- **Alerting and Reporting Engine**: Generates alerts for detected malware and provides reporting capabilities for system administrators and stakeholders.
- **Data Storage**: Stores both raw and processed data for auditing, analysis, and future use.

3. Functional Requirements

3.1 Malware Detection

The system should be able to accurately detect malware in the provided data.

3.2 Malware Classification

The system should categorize detected malware into specific types (e.g., viruses, worms, trojans).

3.3 Alerting and Reporting

The system should generate alerts for detected malware and provide reporting capabilities for system administrators.

4. Non-Functional Requirements

4.1 Performance

The system should be able to process a high volume of data with minimal latency.

4.2 Scalability
The architecture should support horizontal scalability to handle increasing data volumes.

4.3 Reliability and Availability
The system should have a high level of uptime and be resilient to failures.

4.4 Security
The system should implement strong security measures to protect against unauthorized access and data breaches.

4.5 Maintainability
The architecture should allow for easy maintenance, updates, and troubleshooting.

5. Data Flow

5.1 Ingestion
Data is collected from various sources including files, network traffic, and endpoints. This data is then sent to the pre-processing module.

5.2 Processing
The pre-processing module cleans and prepares the data for analysis. This includes normalization and feature extraction.

5.3 Storage
Both raw and processed data are stored in the data storage component for auditing, analysis, and future use.

5.4 Analysis
The machine learning model analyzes the pre-processed data for malware detection and classification.

5.5 Reporting
The alerting and reporting engine generates alerts for detected malware and provides reporting capabilities for system administrators.

6. Technology Stack

6.1 Programming Languages

- Python (for machine learning model and data processing)
- Java (for backend components)

6.2 Frameworks and Libraries

- Scikit-learn (for machine learning)
- Flask (for web application)
- Apache Kafka (for data ingestion)
- Apache Spark (for data processing)

6.3 Databases

- PostgreSQL (for structured data storage)
- MongoDB (for unstructured data storage)

6.4 Operating Systems

- Linux (for server environments)

6.5 Networking

- Virtual Private Cloud (VPC) for network isolation
- SSL/TLS for secure communication

7. Deployment Architecture

7.1 Cloud Provider and Region

- Amazon Web Services (AWS)
- Region: US

7.2 Virtual Machines/Containers

- Docker containers managed by Kubernetes for container orchestration.

7.3 Load Balancing and Autoscaling

- AWS Elastic Load Balancer (ELB) for distributing traffic.
- Autoscaling groups to dynamically adjust compute capacity.

7.4 Data Backup and Recovery

- Automated backups to Amazon S3 for data recovery.

8. Monitoring and Alerting

8.1 Logging and Monitoring Tools

- AWS CloudWatch for monitoring.
- ELK Stack (Elasticsearch, Logstash, Kibana) for centralized logging.

8.2 Key Performance Indicators (KPIs)

- Latency, Throughput, Error Rate, System Uptime.

8.3 Alerting Mechanisms

- Email notifications for critical alerts.
- Slack channel integration for team communication.

9. Security Measures

9.1 Authentication and Authorization

- Role-based access control (RBAC) for user permissions.
- Multi-factor authentication for critical operations.

9.2 Encryption

- Data in transit: SSL/TLS encryption.
- Data at rest: AES-256 encryption.

9.3 Firewalls and Access Controls

- Security groups and Network Access Control Lists (NACLs) for network-level security.
- Web application firewall (WAF) for application-level security.

9.4 Regular Security Audits

- Periodic penetration testing and code reviews.

10. Compliance and Regulations

10.1 Data Privacy Regulations

- GDPR (General Data Protection Regulation)
- CCPA (California Consumer Privacy Act)

10.2 Industry Standards

- ISO/IEC 27001 for Information Security Management
- NIST Cybersecurity Framework

11. Integration with Existing Systems

11.1 Data Sources

- Integration with existing antivirus solutions.
- API endpoints for data ingestion from endpoints.

11.2 APIs and Webhooks

- Provide APIs for integration with internal systems.
- Support webhooks for real-time alerting to other systems.

12. Scalability and Future Considerations

12.1 Scaling Strategies

- Horizontal scaling for increased data volumes.
- Performance tuning for optimization.

12.2 Future Enhancements

- Support for additional machine learning models.
- Integration with threat intelligence feeds for enhanced detection.

13. Dependencies and Third-Party Services

13.1 External Services

- Integration with threat intelligence platforms.
- Utilization of cloud-based services (e.g., AWS services).

13.2 APIs and SDKs

- Utilization of APIs and SDKs for specific functionalities (e.g., AWS SDKs).

14. Cost Estimation

14.1 Cloud Costs

Breakdown of AWS Service Costs:

- **Compute (EC2 Instances):**
    - Monthly cost: $5,000
    - Breakdown:

        - Instance type A: $2,500
        - Instance type B: $2,000
        - Instance type C: $500
- **Storage (S3, EBS):**
- Monthly cost: $2,500
- Breakdown:
    - S3: $1,500
    - EBS: $1,000
- **Networking (Data Transfer):**
    - Monthly cost: $1,000

Cost Optimization Strategies:

- **Reserved Instances:**
    - Reserved Instance Discount: 30%
    - Savings: $1,500/month
    - New Monthly Cost: $3,500
- **Spot Instances:**
    - Spot Instances Usage: 20%
    - Savings: $1,000/month
    - New Monthly Cost: $2,500

14.2 Licensing and Subscription Fees

- **Third-Party Tools:**
    - Antivirus Software: $2,000/month
    - Threat Intelligence Feed: $1,500/month
    - Total: $3,500/month
- **Libraries and Services:**
    - Machine Learning Library License: $1,000/month
    - Database License: $500/month
    - Total: $1,500/month

14.3 Operational Costs

- **Ongoing Maintenance and Monitoring:**
    - Monitoring Tools Subscription: $1,200/month
    - Maintenance Personnel: $3,000/month
    - Total: $4,200/month
- **Employee Training and Support:**
    - Training Programs: $2,500/month
    - Support Contracts: $2,000/month
    - Total: $4,500/month

15. Disaster Recovery and Redundancy

15.1 Backup and Restore Procedures

- Regular backups of critical data and configurations.
- Procedures for restoring services in case of failures.

15.2 Redundancy Measures

- Implementing redundant components for critical systems (e.g., multi-region deployments).