GT-BIS
Mecanismos para Análise de Big Data em Segurança da Informação

EQUIPE

Coordenador
Daniel Macêdo Batista
Instituto de Matemática e Estatística (IME)
Universidade de São Paulo (USP)

Coordenadores-adjuntos
Luiz Arthur Feitosa dos Santos
Rodrigo Campiolo
Diego Bertolini
Universidade Tecnológica Federal do Paraná (UTFPR)

Gerência de Projeto
Wagner A. Monteverde (UTFPR)
Marlon Fernandes Antonio (UTFPR)

Pesquisa e Desenvolvimento
Henrique Sousa Pinheiro (UTFPR)
Vinícius Ribeiro Morais (UTFPR)
Renan Viana Hoshi (UTFPR)
Matheus Sapia Guerra (UTFPR)
Éderson Cássio L. Ferreira (USP)
Erika Guetti Suca (USP)

Parceiros
Universidade de São Paulo (USP)
Universidade Tecnológica Federal do Paraná (UTFPR)

SITE
gtbis.ime.usp.br

CONTATO
gt-bis@listas.rnp.br

DESCRIÇÃO

O protótipo analisa quantidades massivas de dados a fim de detectar incidentes contra a segurança de ambientes computacionais. Esses dados são obtidos a partir do monitoramento de *logs* de diferentes serviços (servidores *web*, SGBD, *syslog*, entre outros) e de sistemas de segurança (IDS, *firewall*, *honeypots*). Durante a análise, são empregadas técnicas de Inteligência Artificial e Aprendizado de Máquina para realizar a correlação de dados e identificação de ciberameaças. Dessa forma, o protótipo é capaz de detectar ataques em redes de computadores e evidenciar ameaças que passariam despercebidas (falsos negativos) por outros sistemas.

Uma arquitetura (Figura 1) foi proposta e implantada para detectar ameaças em grandes volumes de dados de segurança e de serviços de organizações. De forma resumida, os módulos dessa arquitetura e as tecnologias empregadas são:

- Sensores: coletam dados e normalizam logs de serviços e aplicações. No protótipo, há sensores que monitoram os servidores HTTP Apache e Ngnix; *logs* do *syslog*, principalmente para analisar *iptables*, MySQL; *Honeypots* Kippo e Honeywrt; e os IDS Bro, Snort e Suricata.

- Processamento: processam os dados coletados nas redes locais e informações externas. As tecnologias empregadas são:
  - Logstash: gerencia, normaliza e enriquece os logs;
  - Kafka: viabiliza o processamento dos dados em tempo real, isto é, leitura, escrita, armazenamento e manutenção dos fluxos de dados;
  - Spark: processa grandes volumes de dados em tempo real. **Os principais avanços no estado da arte alcançados com o protótipo são os serviços propostos e implementados no Spark**: (a) processamento de reputação: correlaciona informações de reputação internas e externas; (b) processamento de *logs:* processa *logs* e extrai características; (c) aprendizado: gera modelos de aprendizado de máquina em um conjunto de dados de treino; (d) detecção de ameaças: faz uso de modelos para a detecção de ameaças.
  - ElasticSearch: gerencia o armazenamento e recuperação de grandes volumes de dados.

- Visualização e interação: provê a interação do administrador com os componentes da arquitetura e visualização do estado do sistema. Esse módulo atualmente é representado por uma interface *web*.

GT-BIS – Mecanismos para análise de Big Data em segurança da informação

A arquitetura foi implantada por meio de máquinas virtuais instanciadas e replicadas na Universidade Tecnológica Federal do Paraná (UTFPR) e na Universidade de São Paulo (USP). Experimentos avaliaram os diferentes módulos da arquitetura, dos quais se destacam a adição de novos sensores, a criação de novos modelos de detecção, a importação e associação de informações de reputação e a capacidade de escalar horizontalmente e verticalmente. **As taxas de verdadeiros positivos alcançados nos experimentos foram de mais de 98%.** Os desafios para a Fase 2 são treinar novos modelos de detecção e adicionar módulos para compartilhar dados de diferentes organizações e, assim, correlacionar e detectar ciberameaças antecipadamente ou mais rapidamente fazendo uso de infraestrutura em nuvem.

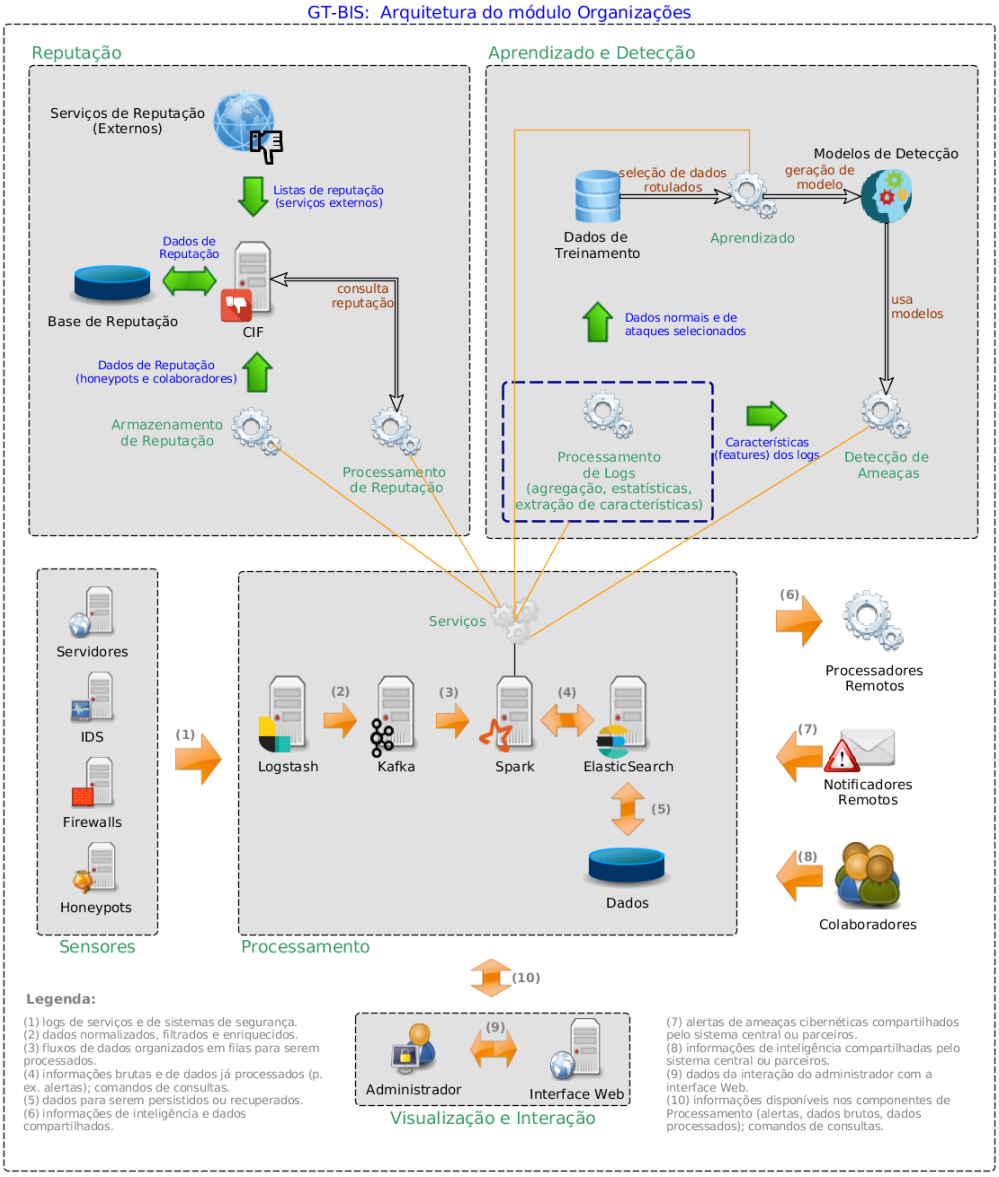

Figura 1