Universidade de Coimbra
Mestrado em Engenharia e Ciência de Dados
Disciplina de Segurança e Privacidade
Ano Letivo 2022/2023

Projeto 1: Data Sharing with Encryption

Introdução

Este projeto teve como objetivo desenvolver um sistema de troca de informação seguro entre duas empresas, *ControlER* e *Delentture.* Neste sistema, a empresa ControlER quer enviar informação que tem sobre os seus clientes à empresa Delentture para que esta possa analisar os dados e detetar possíveis fraudes. Para tal, e como ambas as empresas querem manter os dados seguros e livres de eventuais ataques ou roubos, as empresas decidem enviar os seus dados encriptados.

Este trabalho permitiu, assim, aprofundar o nosso conhecimento nas tecnologias existentes para proteger e partilhar dados de forma segura.

Threat model

O modelo escolhido para ajudar a identificar ameaças, ataques, vulnerabilidades e contramedidas que podem afetar foi o "*Stride mode*l", o mesmo subdividido em seis categorias.

Spoofing identity
Evolve o acesso ilícito e posteriormente usar a informação de autenticação de um utilizador, fazendo-se passar pelo mesmo, ocorre devido a uma falha na propriedade de autenticação.
Exemplo: Alguém roubar a informação de autenticação do empregado da Delentture, acabando por ter acesso à data com informações privadas dos clientes da ControlER.

Tampering
Consiste na modificação dos dados maliciosamente, levando a que propriedade de segurança é violada.
Exemplo: Uma pessoa fora das duas empresas, de alguma forma têm acesso ao dataset e altera a mesma, fazendo com que serviço prestado pela Delentture não tenha o resultado desejado.

Repudiation
Resume-se em não admitir o ato de uma determinada ação sem os restantes conseguirem provar que essa pessoa o fez.
Exemplo: O empregado nega qualquer tipo de ação relacionada com o projeto direcionada à ControlER.

Information Disclosure
Envolve a exposição de informações a indivíduos que não deveriam ter acesso a elas, isto faz com que a propriedade de confidencialidade não seja cumprida.
Exemplo: Informações privadas dos clientes da ControlER estão disponíveis na Internet.

Joana Simões, n.º 2019217013
Tomás Ferreira, n.º 2019224786

Denial of Service
Indisponibilidade dos serviços, o que não permite realizar a propriedade de disponibilidade.
Exemplo: A Delentture não consegue fornecer os serviços à ControlER.

Elevation of Privilege
Um empregado sem privilégios obtém acesso a informações não autorizadas, consequentemente, tem acesso suficiente para comprometer ou destruir todo o sistema. Não respeitando, assim, a propriedade de autorização.
Exemplo: Um empregado da Delentture apagar todos os serviços feitos.

Design do sistema e modelo de comunicação

Key exchange

Para estabelecer uma chave privada partilhada entre ambas as empresas, foi utilizado o algoritmo do *Diffie-Helman* (DH).

Encriptação e desencriptação dos dados

Na encriptação e desencriptação dos dados foram utilizados dois mecanismos diferentes de chave simétrica. Ambos utilizam encriptação de chave simétrica usando AES, mas diferem no modo de aplicação deste. Quando é necessária a autenticidade dos dados, é usada uma encriptação autenticada com o AES-GCM que utiliza o modo *Galois Counter Mode*. Por outro lado, no caso de apenas ser necessária a confidencialidade dos dados, é usado o CounTer Mode (CTR) do AES, por permitir a paralelização da encriptação e desencriptação dos blocos. De notar, que apesar de o CTR permitir a paralelização, esta não é implementada pela biblioteca utilizada.

A escolha de utilizar um algoritmo de encriptação diferente, o AES-GCM, quando é necessária a autenticação e integridade dos dados, derivou de que utilizando este algoritmo não seria necessário gerar uma nova chave privada partilhada entre ambas as empresas. Se fosse utilizado sempre o modo CTR para encriptação combinado com outro algoritmo, como o *Hash-based message authentication* (HMAC), seria necessário partilhar uma nova chave privada utilizando o DH. Assim, como a encriptação autenticada utilizando o AES-GCM garante confidencialidade, autenticidade e integridade sem necessitar de mais nenhuma chave privada, optou–se por ele.

Autenticidade e integridade dos dados

Como já foi referido acima, quando é necessário garantir a autenticidade e integridade dos dados, é utilizada uma encriptação autenticada através do algoritmo AES-GCM.

Joana Simões, n.º 2019217013
Tomás Ferreira, n.º 2019224786

Universidade de Coimbra
Mestrado em Engenharia e Ciência de Dados
Disciplina de Segurança e Privacidade
Ano Letivo 2022/2023

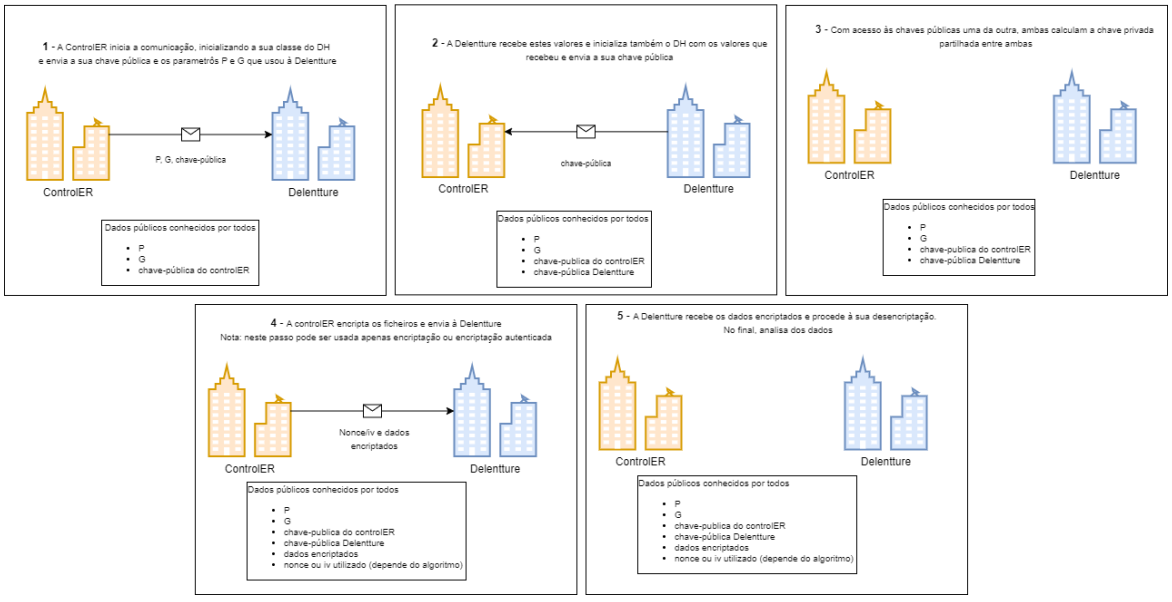
Modelo de comunicação

O processo de comunicação é iniciado pelo ControlER. Este inicializa o Diffie-Hellman (DH) e comunica os valores utilizados (P e G) bem como a sua chave-pública à Delentture. Quando esta recebe os valores, inicializa também o DH e partilha a sua chave-pública. Quando ambas as empresas têm acesso à chave-pública uma da outra, ambas calculam a chave-privada resultante entre ambas.

Com a chave-privada partilhada, os dados são encriptados pela ControlER, podendo, ou não, também ser autenticados antes de serem enviados para a Dellenture. Ao receber o ficheiro encriptado, a segunda empresa inicia o processo de desencriptação e autenticação, no caso desta estar ativa. A seguir, procede à análise dos dados.

Neste projeto, e ao contrário do que é utilizado na realidade, a comunicação não foi realizada por sockets. Para simular a comunicação entre empresas em zonas distintas, foram utilizados ficheiros para a troca de mensagens. Foram utilizados dois tipos de ficheiros: (1) ficheiros binários para as trocas das chaves públicas e dos dados encriptados por estarem em bytes; (2) um ficheiro *json* para valores inteiros como o *P* e *G* usados no DH. Toda a troca de mensagens, ou seja, a escrita em ficheiros, está auxiliada por uma classe *Communication* a que ambos têm acesso.

A figura abaixo mostra um diagrama a sumarizar de todo o processo de troca de informação entre as empresas (para observar melhor, abrir o documento diagrama-SP.png).



Joana Simões, n.º 2019217013
Tomás Ferreira, n.º 2019224786

Universidade de Coimbra
Mestrado em Engenharia e Ciência de Dados
Disciplina de Segurança e Privacidade
Ano Letivo 2022/2023

Análise dos resultados

	ControlER	Dellenture
DH - Cálculo da chave privada e pública	29s	6.35ms
DH - Cálculo da chave privada partilhada	7.81ms	6.01ms

Ao analisar os tempos de execução do DH, pode-se retirar que o tempo de calcular a chave pública e privada no ControlER é muito superior ao tempo da Delentture. Esta diferença deve-se ao facto de no lado do ControlER, também incluir o cálculo dos valores de P e G comuns a ambas as empresas, sendo este um processo demorado.

	AES-CTR	AES-GCM
Encriptação do Infringement_dataset	339ms	226ms
Encriptação do Columns_description	3ms	3.06ms
Desencriptação do Infringement_dataset	2.89s	2.78s
Desencriptação do Columns_description	5.45ms	5.85ms

Por observação da tabela anterior, podemos concluir que os processos de encriptação e desencriptação sem autenticidade dos dados apresentam tempos de execução superiores em comparação com os processos homólogos que acrescentam uma camada de autenticidade.

Em teoria, a encriptação autenticada deveria demorar mais tempo. No entanto, apesar de o algoritmo o permitir, a biblioteca utilizada neste projeto, não implementa a paralelização da encriptação de cada bloco. Além disso, o algoritmo AES-GCM é mais utilizado atualmente, estando otimizado para usar o paralelismo, justificando assim a diferença de tempos.

É ainda de realçar que, como esperado, existem diferenças significativas nos tempos de encriptação/desencriptação consoante o tamanho dos ficheiros utilizados. Estas diferenças verificam-se em ambos os algoritmos usados, mostrando que o tamanho do ficheiro afeta bastante o tempo de encriptação/desencriptação dele.

Joana Simões, n.º 2019217013
Tomás Ferreira, n.º 2019224786

Por último, é também de destacar que a desencriptação dos ficheiros demora muito mais tempo que a encriptação deles. Sendo que este valor se acentua com o aumento do tamanho do ficheiro.

Discussão

Na criação deste sistema foram discutidas diversas abordagens, especialmente no que toca a que algoritmos usar em cada fase do projeto.

Pela visualização dos tempos de execução obtidos, é possível ver o impacto que o tamanho do ficheiro e o algoritmos usados têm nos tempos finais. Vemos ainda que gerar valores aleatórios iniciais de P e G para o *Diffie-Hellman* demora um tempo bastante significativo. Conclui-se, assim, que todas as escolhas de implementação afetam a performance, tanto em segurança, como em tempo, de todo o processo de comunicação entre empresas, sendo importante refletir bem elas.

Por último, é importante referir que numa análise inicial dos dados, verificou-se que alguns parâmetros poderiam ser codificados antes de serem enviados à Delentture, uma vez que incluem informação privada dos clientes. No entanto, optou-se por não os codificar ou não enviar por estar foram do âmbito do projeto, ficando para um trabalho futuro. Outro trabalho fora do âmbito do projeto seria enviar os dados no sentido contrário. Após a Delentture analisar os dados, faria um relatório que enviaria à ControlER, também este encriptado uma vez que poderia conter padrões interessantes sobre os clientes e que a ControlER não quer revelar a outras empresas concorrentes. No entanto, a abordagem seria a mesma que foi utilizada, ou seja, os algoritmos e chaves manter-se-iam, mudando apenas o sentido da comunicação.

Conclusão

Com este projeto foi possível colocar em prática os conhecimentos adquiridos nas aulas, e perceber de uma forma mais prática o impacto que diferentes algoritmos podem ter na partilha de dados entre empresas, embora que em pequena escala no nosso caso. O desenvolvimento deste trabalho deu-nos ainda uma perceção de como realmente ocorre a interação entre empresas, e quais as medidas de precaução e segurança que devemos ter a fazer o nosso trabalho de *data scientists*.

Referências

AES — PyCryptodome 3.15.0 documentation. (n.d.). Retrieved October 13, 2022 from PyCryptodome's documentation: https://pycryptodome.readthedocs.io/en/latest/src/cipher/aes.html

Joana Simões, n.º 2019217013
Tomás Ferreira, n.º 2019224786

Authenticated encryption. (n.d.). Retrieved October 13, 2022 from Authenticated encryption — Cryptography 39.0.0.dev1 documentation: https://cryptography.io/en/latest/hazmat/primitives/aead/#cryptography.hazmat.primitives.ciphers.aead.AESGCM

Classic modes of operation for symmetric block ciphers — PyCryptodome 3.15.0 documentation. (n.d.). Retrieved October 13, 2022 from PyCryptodome's documentation: https://pycryptodome.readthedocs.io/en/latest/src/cipher/classic.html#ctr-mode

Diffie-Hellman key exchange. (n.d.). Retrieved October 13, 2022 from Diffie-Hellman key exchange — Cryptography 39.0.0.dev1 documentation: https://cryptography.io/en/latest/hazmat/primitives/asymmetric/dh/

Hash-based message authentication codes. (n.d.). Retrieved October 13, 2022 from Hash-based message authentication codes (HMAC) — Cryptography 39.0.0.dev1 documentation: https://cryptography.io/en/latest/hazmat/primitives/mac/hmac/

Key Serialization. (n.d.). Retrieved October 13, 2022 from Key Serialization — Cryptography 39.0.0.dev1 documentation: https://cryptography.io/en/latest/hazmat/primitives/asymmetric/serialization/

Joana Simões, n.º 2019217013
Tomás Ferreira, n.º 2019224786