Gestão e Pesquisa de Dados Privados em Nuvens de Armazenamento

Bernardo Ferreira, Henrique Domingos

Departamento de Informática, FCT/UNL
CITI Centro de Informática e Tecnologias da Informação
bernardolferreira@gmail.com,hj@fct.unl.pt

Resumo. O artigo apresenta uma solução para armazenamento e gestão de dados privados mantidos em nuvens de armazenamento na internet (ou I*nternet Storage Clouds*). A solução suporta operações sobre os dados mantidos cifrados, incluindo leituras, escritas e pesquisas baseadas em multi-palavras e classificação de relevância dos mesmos. A aproximação baseia-se numa arquitetura de sistema *middleware* que utiliza técnicas de encriptação homomórfica combinadas com mecanismos de indexação dinâmica, preservando condições de privacidade, sem necessidade de se decifrarem os dados durante as operações na C*loud* e sem transferência dos mesmos durante as pesquisas.

Palavras-Chave: *Nuvens de Armazenamento de Dados na Internet, Segurança e Privacidade, Cifras Homomórficas, Pesquisa e Classificação de Documentos.*

1 Introdução

O acesso indevido ou a divulgação não autorizada de dados privados mantidos em *Clouds* de armazenamento tem sido referido como um problema crítico, não apenas na salvaguarda segura dos dados mas também na preservação de condições de segurança de dados acedidos por aplicações "*on-line*" [1]. A dependência das *Clouds* de bases de confiança de terceiros impede o controlo ou auditoria integral pelos utilizadores sobre eventuais vulnerabilidades de segurança na infraestrutura computacional (hardware/software) dos provedores, podendo os dados ser objeto de ações ilícitas ou operação descuidada de técnicos ou administradores dos servidores [2].

No presente artigo propõe-se uma solução que tem em vista a conjugação de requisitos de segurança, privacidade, fiabilidade e disponibilidade permanente dos dados, sob controlo independente do utilizador, de forma a tornar confiáveis as soluções de armazenamento e gestão de dados em *Clouds* de armazenamento. A conjugação das anteriores dimensões é endereçada por uma arquitetura de referência concebida como um sistema *middleware* para intermediação de serviços de armazenamento seguro e pesquisa de dados privados mantidos em *Clouds* de armazenamento. A contribuição do artigo foca-se no suporte de pesquisas seguras sobre os dados, propondo-se mecanismos efetivos para pesquisa por relevância de documentos, com múltiplas palavras-chave, e acesso à informação cifrada na *Cloud* mantendo condições de privacidade dos dados e pesquisas sob total controlo dos utilizadores. A abordagem utiliza esquemas criptográficos que exploram técnicas de encriptação homomórfica [3] combina-

das com mecanismos de indexação dinâmica [4]. Esta combinação permite preservar condições de privacidade em diferentes cenários de implementação, sem necessidade de se decifrarem os dados ou de proceder à sua transferência durante as pesquisas. A gestão dos dados privados pode assim ser feita sob total controlo dos utilizadores, independentemente do provedor.

2 Modelo de sistema e arquitetura da solução proposta

A solução proposta visa conjugar diferentes requisitos num modelo e arquitetura de sistema (*middleware*), para gestão de privacidade, autenticidade, fiabilidade e disponibilidade permanente de dados guardados em *Clouds* de armazenamento na Internet, sob o controlo dos utilizadores que os possuem. Assim, uma solução para o problema deve endereçar os seguintes requisitos:

- Garantir a confidencialidade, integridade e privacidade de dados, com controlo e auditoria independente dos utilizadores e estendendo estas garantias às operações de pesquisa sobre os mesmos, nomeadamente suportando pesquisas por multi-palavras chave e ordenação por relevância;
- Permitir a viabilidade da solução para integração com serviços de C*louds* de armazenamento de provedores de Internet, aproveitando todas as vantagens que estes possam oferecer;
- Complementar as soluções de segurança e privacidade dos dados e operações com requisitos adicionais de fiabilidade e disponibilidade permanente, sem que tais garantias estejam dependentes de cada provedor em particular;
- Conceber uma solução cuja operação possa ser totalmente controlada pelos utilizadores, permitindo condições de independência em relação aos serviços de diferentes provedores, o que permitirá defender contra práticas comerciais que possam ser lesivas dos seus interesses (como por exemplo, práticas de *vendor lock-in* [5]).

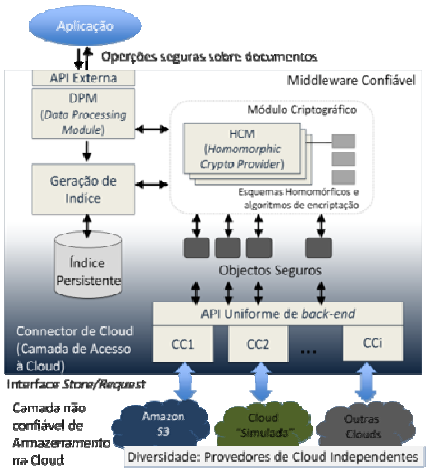

Fig. 1. Arquitetura de referência da solução perspetivada

Complementarmente, pretende-se uma solução que permita flexibilizar a sua implementação para diferentes opções de suporte dos serviços *middleware*, incluindo os seguintes casos de uso: **(1)** *Middleware* instalado no dispositivo ou computador pessoal do utilizador, sendo este uma base computacional de confiança; **(2)** *Middleware* seguro (constituindo uma base computacional de confiança) a funcionar como serviço proxy numa rede local; **(3)** *Middleware* como serviço na *Cloud*, desde que certas condições do modelo de ataques não se verifiquem.

A figura 1 mostra a arquitetura de referência subjacente à abordagem visionada para a solução. A arquitetura tem em vista o uso de técnicas de encriptação homomórfica (módulo HCM) conjuntamente com processos de indexação dinâmica dos dados. Como se representa, perspetiva-se a exploração da diversidade natural de diferentes *Clouds* de armazenamento, permitindo melhorar a fiabilidade, segurança e disponibilidade da solução.

3 Implementação e avaliação

De modo a validar experimentalmente a solução concebida, foi implementado um protótipo do sistema *middleware* e foram realizados diferentes testes de carga. Os testes realizados dividiram-se em dois grupos e, em ambos os casos, foram suportados pelo serviço de *Cloud* Amazon S3. O primeiro grupo de testes validou o desempenho de escrita e indexação da coleção em diferentes cenários de implementação do middleware, comparando-se com a escrita direta dos documentos na *Cloud* (através da solução segura da própria Amazon para cifra de dados no lado do cliente). O segundo grupo de testes comparou a performance de pesquisas nos diferentes cenários de implementação e execução da solução.

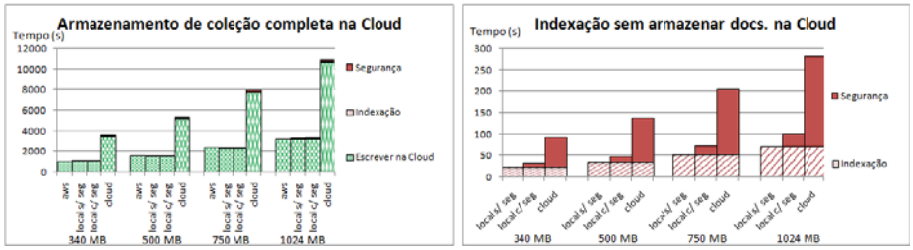

Fig. 2. Performance de escrita (2(a)) e indexação (2(b)) nos diferentes cenários

O gráfico 2(a) representa os tempos de escrita completa da coleção de RFCs do IETF, incluindo indexação e processos de segurança. O gráfico 2(b) representa apenas os tempos de indexação e segurança para cada cenário, sem escrita na *Cloud*. Os resultados mostram que o *overhead* introduzido pelos processos do *middleware* é mínimo face ao custo de enviar para a *Cloud* a mesma coleção de documentos. O caso de uso do *middleware* como serviço na *Cloud* apresenta piores resultados de performance, incluindo escrita de mais informação na *Cloud*, devido ao custo adicional de garantir privacidade e segurança dos dados, sob controlo dos utilizadores, num ambiente não controlável pelos mesmos.

A figura 3 mostra o resultado do segundo grupo de testes. Em todos os cenários de implementação do sistema conseguem-se tempos de pesquisa abaixo dos 500 milissegundos, o que demonstra a viabilidade da solução (sempre inferior a 1 segundo, sendo adequado a aplicações "*on-line*"). Os dados da figura 3 foram obtidos usando uma coleção de 340MB. No entanto, com a conjugação de estruturas de dados eficientes e adequadas à solução com técnicas de *Information Retrieval* [4], é possível garantir a performance descrita independentemente do tamanho da coleção.

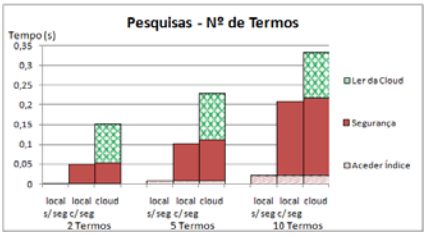

Fig. 3. Performance de pesquisas nos diferentes cenários

4 Conclusões

O artigo apresenta uma solução que tem em vista a conjugação de requisitos de fiabilidade, disponibilidade, segurança e privacidade de dados mantidos em *Clouds* de armazenamento de provedores de Internet. A solução é endereçada como um sistema *middleware* para intermediação de serviços de armazenamento seguro nas *Clouds*. Este sistema suporta a gestão e armazenamento de dados privados, sob total controlo do utilizador e independentemente de diferentes *Clouds* que possam ser utilizadas. Uma contribuição relevante da solução proposta foca-se no suporte de pesquisas por multi-palavras chave, com critérios de classificação de documentos usando métricas de relevância e *ranking* dos dados. Durante as operações de pesquisa preservam-se condições de privacidade, sob total controlo dos utilizadores. A abordagem apresenta esquemas criptográficos que exploram técnicas de encriptação homomórfica combinadas com mecanismos de indexação dinâmica. A implementação do sistema proposto e a sua avaliação mostram que a solução é viável, oferece mais segurança e maior controlo do utilizador (comparativamente a uma solução promovida pela Amazon AWS) e não agrava condições de latência de acesso e de disponibilidade dos dados.

Referências

1. Privacy Rights Clearinghouse. Chronology of data breaches. http://www.privacyrights.org/data-breach.
2. A. Chen. GCreep: Google engineer stalked teens, spied on chats. Gawker, September 2010. http://gawker.com/5637234/.
3. R. Popa, C. Redfield, N. Zeldovich, H. Balakrishnan. *CryptDB: Protecting Confidentiality with Encrypted Query Processing*. SOSP '11, October 23–26, 2011, Cascais, Portugal.
4. C. Manning, P. Raghavan, H. Schütze. "An Introduction to Information Retrieval", Cambridge University Press, 2009
5. A. Bessani, M. Correia, B. Quaresma, F. André, P. Sousa. *DEPSKY: Dependable and Secure Storage in a Cloud-of-Clouds*. EuroSys'11, April 10–13, 2011, Salzburg, Austria