Caixas de Interesses: um Novo Mecanismo para a Colaboração através de Nuvem de Armazenamento de Dados

Felipe Leite da Silva, Roberto Araújo, Lucas Melo Silva, Nelson Neto Sampaio
Laboratório de Segurança e Criptografia Aplicada (LabSC)
Faculdade de Computação
Universidade Federal do Pará (UFPA)
Pará, Brasil
Email:{fsilva,rsa,lucasmelo,nelsonneto}@ufpa.br

ABSTRACT
Individuals and organization have increasingly adopted cloud computing services for outsourcing computational resources. Data storage, specifically, presents one of the most popular outsourced resources. Using cloud storage services, clients meet their needs obtaining a data repository for archive and backup files. Usually, these services offer unlimited storage capacity along with a high availability and easy multi-platform access infrastructure. In addition, they also offer collaboration capability between users. The data sharing mechanisms provide the primary means of collaboration and allow users to work from remote locations and exchange large amounts of information.
Although traditional mechanisms present benefits, they also restrict file sharing among users. By design, they limit sharing capability to group of users that know each other. In particular, cloud storage services do not fully support content share between users with no social ties, but with common interest (e.g., researchers with the same area of expertise).
Targeting this scenario, we introduce a sharing mechanism based on so-called *interest box* as a novel approach for sharing content in cloud storage services. This mechanism enable file sharing between users with a potential common interest based on their stored content. Moreover, the user controls who can access his files using attributes from an attribute provider center, allowing only authorized access to his data.

Keywords
Cloud, sharing files, cloud storage collaboration.

1. INTRODUÇÃO
O surgimento de novas tecnologias e o aprimoramento dos dispositivos móveis impulsiona um amplo crescimento da quantidade de dados produzidos tanto por organizações quanto por pessoas. Mediante a crescente necessidade de armazenamento desses dados, diversos serviços surgem a fim de atender a tal demanda.

Nesse contexto, as nuvens de armazenamento destacam-se como uma das principais soluções para o armazenamento de dados. Esses serviços oferecem aos seus usuários capacidade de armazenamento de acordo com suas necessidades e incluem outros benefícios, como a alta disponibilidade (ex. arquivos são acessíveis a qualquer momento) e o acesso multiplataforma por meio de diferentes dispositivos [1][2].

Além disso, essas nuvens caracterizam-se por facilitar a colaboração entre usuários. O compartilhamento de dados, em particular, é um dos principais recursos ofertados para promover tal colaboração. Por meio dele, os usuários podem trocar grande quantidade dados entre si e trabalhar de forma integrada. Por exemplo, organizações podem compartilhar dados dinamicamente entre suas filiais e pesquisadores de institutos diferentes podem cooperar em trabalhos multinstitucionais.

Tradicionalmente, os mecanismos de compartilhamento presentes em nuvens de armazenamento de dados têm seu funcionamento baseado na delegação explícita de acesso aos arquivos. Em outras palavras, o usuário determina regras de acesso para seus arquivos definindo outros usuários do serviço que poderão acessá-los. Essas regras são definidas através da utilização de um identificador (ex. e-mail, nome de usuário) referente ao receptor do arquivo compartilhado. Exemplos de mecanismos desse tipo são as listas de controle de Acesso (*Access Control List - ACL*) como as utilizadas pela Amazon S3 [3] e os níveis hierárquicos (*Role Based Access Control - RBAC*) como utilizados pela nuvem Rackspace [4]. Esta forma de compartilhamento possibilita a colaboração dinâmica e controlada entre os usuários, pois eles podem definir a qualquer momento quem terá acesso aos seus arquivos.

Apesar dos benefícios, esses mecanismos apresentam uma limitação inerente quando utilizados juntamente com nuvens de armazenamento: eles requerem o conhecimento prévio do identificador do usuário receptor do arquivo compartilhado. Tendo em vista que nuvens de armazenamento são serviços que agregam cada vez mais usuários, esses mecanismos apenas viabilizam o compartilhamento entre conjuntos de usuários que de alguma forma obtiveram os identificadores de outros (ex. amigos, familiares, trabalhadores de uma empresa). Em particular, usuários que não se conhecem, mas que poderiam se beneficiar compartilhando dados entre si (ex. pesquisadores de uma mesma área de atuação que não se conhecem) não são comtemplados por estes mecanismos.

Por outro lado, os serviços de armazenamento em nuvem possibilitam a criação de arquivos ou pastas públicas como alternativa a delegação explícita de acesso a arquivos. Essa forma de compartilhamento baseia-se na remoção de restrições de acesso sobre o item selecionado tornando-o disponível ao público em

geral. Tal solução possibilita um compartilhamento mais abrangente, pois qualquer pessoa poder acessar o arquivo publicado. No entanto, ela compromete o controle do usuário sobre seus dados, uma vez que ela não oferece uma forma de definir restrições de acesso ao arquivo. Outra limitação está relacionada ao suporte para divulgação dos arquivos publicados. Normalmente, os dados são disponibilizados por meio de um *link* público. Esse *link*, por sua vez, deve ser divulgado através de uma plataforma externa (ex. sites pessoais) dificultando a obtenção do arquivo por outros usuários.

Para fim de exemplificação, o cenário a seguir captura as limitações relativas ao compartilhamento em nuvens de armazenamento de dados. Duas pessoas (Alice e Bob) são usuários de um mesmo serviço de armazenamento de dados em nuvem. Este serviço consiste em uma nuvem comunitária em que pesquisadores de todo o país armazenam informações sobre suas pesquisas. Alice é uma usuária que armazena, na nuvem, documentos referentes a sua pesquisa. Embora alguns documentos de Alice devam ser mantidos de forma privada (ex. resultados parciais de pesquisas não publicadas) outros não comprometeriam sua privacidade caso fossem acessados por terceiros (ex. artigos e outros documentos referentes a pesquisas publicadas). Bob, por sua vez, é um pesquisador da mesma área de conhecimento que Alice, mas que não a conhece. Nesse contexto, Alice e Bob são usuários que poderiam se beneficiar através do compartilhamento de dados, mas que não são completamente contemplados pelos mecanismos disponíveis. Visto que não se conhecem, tanto Alice quanto Bob não trocariam identificadores entre si. Além disso, o compartilhamento entre esses usuários pode não ser efetivado caso criem *links* públicos para seus arquivos. Isso porque esta forma de compartilhamento não provê meios de divulgação do arquivo publicado a outros usuários da nuvem. Ele apenas remove as restrições de acesso sobre os dados.

Mediante as limitações destacadas, este artigo apresenta um novo mecanismo para o compartilhamento de dados em nuvens de armazenamento. Denominado de compartilhamento por caixas de interesses, ele atua de forma complementar as estratégias tradicionais de compartilhamento e proporciona uma nova forma de colaboração entre usuário do serviço de nuvem. Especificamente, esse mecanismo possibilita que usuários desconhecidos com algum interesse em comum troquem arquivos entre si. Para promover tal compartilhamento, o mecanismo utiliza a tecnologia de deduplicação de dados do serviço de nuvem. Além disso, cada usuário é capaz de controlar o acesso aos seus dados através da utilização de atributos pertencentes a um centro provedor de atributos. Esse provedor é implantado através de serviços de federação de identidade.

Diferentemente de outras abordagens de compartilhamento de dados, como as redes sociais gerais [5] ou profissionais [6][7], esta proposta permite que os usuários usufruam do recurso de armazenamento disponibilizado pelos serviços de nuvem estendendo sua capacidade colaborativa. Além disso, a aplicabilidade do mecanismo proposto não se restringe ao contexto de compartilhamento de dados em nuvens científicas. Ela pode ser estendida para outros tipos de nuvens, permitindo o compartilhamento de dados em outros contextos, como por exemplo, entre pessoas de organizações distintas em uma federação.

Abaixo são listadas as principais contribuições introduzidas por este artigo:

1- Propor um mecanismo de compartilhamento de arquivos que estende a capacidade de colaboração entre usuários de nuvens de armazenamento de dados através do compartilhamento entre usuários que não se conhecem.

2- Propor a utilização da deduplicação de dados como forma de identificação de usuários desconhecidos que apresentam possíveis interesses em comum em nuvens de armazenamento. Essa abordagem usufrui de uma técnica que se difundiu nas nuvens de armazenamento, e auxilia no aprimoramento da capacidade colaborativa dessas plataformas através do mecanismo abordado pelo artigo.

3- Propor um esquema de controle de acesso que permita o compartilhamento controlado em nuvens de armazenamento entre usuários desconhecidos.

4- Apresentar uma discussão inicial sobre o mecanismo proposto.

A organização deste artigo esta conforme descrito a seguir: na Seção 2 são apresentados os trabalhos relacionados ao tema abordado. Em seguida destaca-se na Seção 3 o conceito de deduplicação de dados. Ele servirá de base para o entendimento do mecanismo proposto neste artigo. Então, na Seção 4 o mecanismo é apresentado e descrito e na Seção 5 é realizada uma discussão inicial sobre ele. Por fim na Seção 6, as considerações finais e o direcionamento para trabalhos futuros são apresentados.

2. TRABALHOS RELACIONADOS

No contexto comercial, o compartilhamento de dados tornou-se uma das principais estratégias para viabilizar a colaboração entre usuários de nuvens de armazenamento de dados. Sendo amplamente adotado por diversos serviços[3][4][8], em alguns casos, ele é aprimorado viabilizando a edição simultânea e em tempo real de alguns formatos de arquivos[9][10]. Por outro lado, os mecanismos disponibilizados limitam o compartilhamento, pois baseiam-se no modelo de delegação explicita de acesso e de publicação de arquivos conforme introduzidos na seção anterior.

No contexto científico, algumas propostas visam à utilização colaborativa das nuvens de armazenamento de dados. No Brasil, o grupo de trabalho Computação em Nuvem para a Ciência (GT-CNC)[11][12] realiza trabalhos nesse contexto. Eles propõem o desenvolvimento de uma nuvem de armazenamento cientifica em que os usuários das instituições pertencentes a Comunidade Acadêmica Federada (CAFe) da Rede Nacional de Pesquisa (RNP) possam usufruir do serviço de forma colaborativa compartilhando dados entre si. Além desse grupo, Koulousis *et. al*. [13] também considera a criação de federações de nuvens para promover o compartilhamento de dados científicos no contexto do projeto VPH-Share. Essas propostas têm por objetivo principal unificar recursos de diferentes infraestruturas e assim permitir que os usuários da federação tenham acesso a eles. Contudo, esses trabalhos não apresentam nenhum mecanismo que contemplem o compartilhamento de dados entre usuário com interesses em comum que não se conhecem.

Outras propostas destinam-se a estender a capacidade de colaboração a um escopo mais abrangente de usuários. Chard *et. al*.[14] e outros pesquisadores [15][16][17] apresentam o conceito de nuvens sociais. Elas correspondem a um modelo em que recursos de nuvens computacionais podem ser compartilhados com base nos relacionamentos que seus usuários possuem em redes sociais. Desta forma, esse modelo fundamenta-se na confiança que um usuário possui nas pessoas com que se relaciona

e assim promove a troca de informações de forma mais aberta, pois considera todos os laços da rede social para o compartilhamento.

No entanto, esse tipo de compartilhamento ainda se limita a um grupo restrito de usuários. Especificamente, mesmo ampliando a capacidade de compartilhamento por meio de laços existentes nas redes sociais, os usuários que não se conhecem, mas que poderiam compartilhar dados de interesses em comum entre si não são compreendidos nesse modelo.

Christin *et al* [18] apresenta um mecanismo de controle de acesso que amplia a capacidade de compartilhamento compreendendo usuários desconhecidos. O mecanismo denominado de bolhas privadas baseia se em parâmetros espaço-temporais e permitem, através do uso de dispositivos móveis, o compartilhamento de dados entre usuários que estejam em um mesmo local, mas que não possuem laços sociais. Apesar de não prever o uso de nuvens de armazenamento em seu trabalho, a proposta pode ser estendida para esse contexto. Em contra partida, ela está restrita ao compartilhamento através do uso de dispositivos móveis.

3. DEDUPLICAÇÃO

Em nuvens de armazenamento de dados é comum o acumulo de vários arquivos iguais na infraestrutura do serviço. Por exemplo, dois arquivos de apresentações digitais contendo os mesmos slides. No entanto, ao armazenar o mesmo dado, a nuvem desperdiça espaço. De forma a resolver tal problema, as nuvens utilizam-se de técnicas de deduplicação de dados. Tais técnicas identificam múltiplas cópias de um arquivo e armazenam apenas uma única instância no repositório de dados[19][20].

Comparada com a compressão, a deduplicação é um mecanismo ainda mais atrativo para grandes repositórios. Isso porque enquanto as técnicas de compressão apenas aperfeiçoam o armazenamento tratando a redundância de informações dentro de um arquivo (ex. reduzindo a quantidade de bits para representar um dado), na deduplicação esse tratamento pode ocorrer tanto dentro de um arquivo quanto entre arquivos diferentes com conteúdo igual ou similar (ex. salvando uma única cópia de um fragmento repetido em um ou mais arquivo). Desta forma elevadas taxas de economia dos recursos de armazenamento podem ser obtidas através da deduplicação[21][22]. Mediante esse benefício, ela é uma técnica bastante difundida entre os serviços de armazenamentos de dados em nuvem sendo utilizada por serviços como Dropbox [8] e outros[23][24].

O funcionamento de um mecanismo de duplicação pode ser sintetizado conforme os passos a seguir:

1- Inicialmente é calculado o valor *hash* do arquivo que será enviado. Esse valor representa um identificador único do arquivo e é gerado com base no seu conteúdo através de funções de *hash* criptográfico [25].

2- Esse valor é comparado com outros valores *hash*s referentes aos arquivos armazenados pelo serviço de armazenamento. Esta comparação pode ser realizada tanto pelo cliente do serviço (*source based deduplication*) quanto pelo provedor do serviço (*target based deduplication)*.

3- Quando uma cópia é encontrada, o usuário não envia o arquivo. Ele apenas é referenciado no arquivo armazenada no servidor evitando o armazenamento de dados duplicados.

Por outro lado, as técnicas de deduplicação podem ser aplicadas de forma diferentes conforme a necessidade do serviço de armazenamento. Uma das principais distinções entre as técnicas está relacionada com a granularidade da comparação dos dados. Nesse contexto, a deduplicação pode ser categorizada em duas estratégias principais: a deduplicação em nível de arquivo e deduplicação em nível de blocos.

Na deduplicação em nível de arquivo considera-se todo o arquivo para identificar dados redundantes. Isto é, o valor *hash* é calculado com base em todo o conteúdo do arquivo. Neste caso, não é possível a identificação de arquivos que apenas partes similares, mas apenas de arquivos completamente iguais.

Na deduplicação em nível de blocos, por sua vez, o arquivo é fragmentado antes de iniciar a verificação de redundância. Então, cada fragmento tem seu valor *hash* calculado e comparado com os fragmentos armazenados pelo serviço de armazenamento. Esta estratégia possui maior eficiência na detecção de dados iguais, visto que blocos iguais entre dois arquivos que apresentam pequenas diferenças são detectados, deduplicados e apenas os blocos diferentes do novo arquivo são enviados ao repositório de armazenamento.

4. O NOVO MECANISMO DE COMPARTILHAMENTO DE DADOS EM NUVEM

Conforme apresentado inicialmente, este artigo propõe um novo mecanismo de compartilhamento de dados para nuvens de armazenamento. Nesta seção esta solução é detalhada. Inicialmente, é introduzida uma visão geral do mecanismo na subseção 4.1. Na subseção seguinte são apresentadas definições preliminares a fim de oferecer suporte ao entendimento do mecanismo. Por fim, o funcionamento do mecanismo é descrito na Subseção 4.3.

4.1 Visão Geral

O compartilhamento por caixas de interesses consiste em um mecanismo que permite usuários desconhecidos pertencentes a um mesmo serviço de armazenamento de dados em nuvem compartilharem arquivos de interesses em comum entre si. Esse mecanismo baseia-se na utilização de deduplicação de dados para a identificação de interesses em comum entre usuários. Os usuários armazenam seus dados em estruturas denominadas de caixas de interesses e então os arquivos deduplicados nestas estruturas indicam que outros usuários possuem os mesmos arquivos. Esses usuários são, então, identificados como suscetíveis ao compartilhamento. A partir de então os demais arquivos das caixas de interesses que não foram deduplicados são sugeridos para o compartilhamento de dados.

O mecanismo é realizado por meio da interação entre três elementos principais: uma nuvem de armazenamento, um centro de distribuição de atributos e o conjunto de usuários do serviço de nuvem.

A nuvem de armazenamento é o serviço que os usuários utilizam, através de um cliente de acesso, para armazenar, compartilhar e resgatar arquivos. O centro de distribuição de atributos consiste em uma federação de identidade. Essas federações são formadas a partir de uniões de organizações que desejam compartilhar serviços entre seus usuários, como por exemplo, serviços em nuvem. As federações de identidade gerenciam os usuários

através de provedores de identidade (*Identity Provider - IdP*). Os IdPs disponibilizam as informações de identidade dos usuários aos provedores de serviços da federação (*Service Provider - SP*). Os *SPs*, por sua vez, são os consumidores das credencias. Eles permitem o acesso ao seu serviço mediante credencias válidas. No contexto deste trabalho a nuvem de armazenamento é o serviço que consome as credenciais disponibilizadas pelos IdPs e, portanto, ela representa o *SP* da federação de identidade.

A Figura 1 apresenta estes componentes assim como uma visão geral do fluxo de interações entre eles.

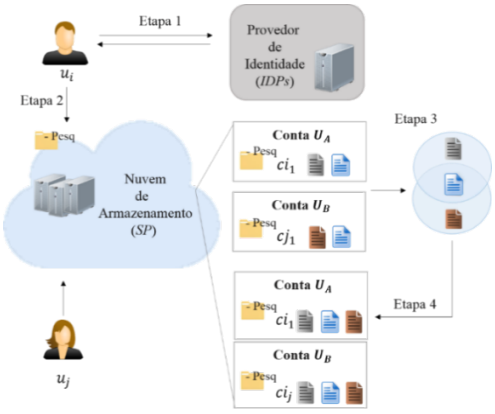

Figura 1. Visão Geral do Mecanismo de Compartilhamento de Dados por Caixas de Interesses.

Conforme se pode observar, as interações ocorrem através de quatro etapas principais. Essas etapas correspondem ao funcionamento do mecanismo de compartilhamento por caixas de interesses.

Na primeira etapa o usuário realiza a autenticação e a obtenção de seus atributos. Neste artigo considera-se que a federação prove o serviço de identidade por meio do protocolo SAML 2.0 (*Security Assertion Markup Language*) [26]. Este protocolo oferece tanto a disponibilização de atributos dos usuários quanto a capacidade de autenticação única entre domínios diferentes (*Single Sing On - SSO*). Ele representa um dos principais protocolos suportado pelas soluções de identidade federada [27][28]. Após esta etapa o usuário possui acesso a nuvem de armazenamento de dados.

Na segunda etapa o usuário cria sua caixa de interesse. Concretamente, essas estruturas são representadas por pastas ou elementos semelhantes existentes em nuvens de armazenamento de dados (ex. *buckets*, *contêineres*). As caixas de interesses se distinguem das pastas através da adição de metadados gerais e específicos que as caracterizam. Os metadados gerais definem uma pasta como uma caixa de interesse e também podem agregar outras informações a ela, como por exemplo, a data de criação e uma descrição. Os metadados específicos são os atributos que o usuário possui na federação. Eles são obtidos na primeira etapa quando o usuário e interage com o serviço de identidade federada.

A fim de promover o compartilhamento, o usuário dever adicionar arquivos na caixa de interesse. Ao adicioná-los, eles se tornam disponíveis para que outros usuários possam obtê-los quando identificados como potenciais receptores. Para controlar a disponibilização dos dados compartilhados, o usuário deve incluir os metadados específicos da caixa de interesse. Desta forma, apenas caixas de interesses com atributos iguais aos da caixa do detentor do arquivo receberão os dados. Além disso, o usuário pode tornar o compartilhamento indisponível removendo o arquivo da caixa de interesse.

Na terceira etapa, o serviço de nuvem atua na identificação de possíveis usuários receptores do compartilhamento. Para isso, ela conta com um componente denominado de Gerenciador de Caixas de Interesse. Esse componente é responsável por encontrar outros usuários que possuam caixas de interesse compatíveis com a caixa do arquivo adicionado na etapa 2.

Por fim, a nuvem promove o compartilhamento dos arquivos na etapa 4. Ela obtém informações de caixas de interesses em comum interagindo com o componente gerenciador. Em seguida, ela sugere o compartilhamento entre os usuários com base nas informações obtidas. O compartilhamento, por sua vez, apenas será efetivado mediante a aceitação do usuário receptor. Desta forma, o receptor pode controlar quais arquivos ele recebe pelo compartilhamento.

4.2. Definições

Nesta seção são introduzidas as principais notações e definições utilizadas na descrição do mecanismo na Seção 4.3.

Definição 1. (Nuvens de Armazenamento de Dados) Uma Nuvem de Armazenamento é uma 4-upla *CSP* = <*U*, *F*, *P*, *R*>. Em que:

- U é o conjunto de todos os usuários u_i da nuvem, $U = \{u_1, u_2, \dots, u_n\}, (1 \leq i \leq n)$.

- F é o conjunto de todos os arquivos f_i armazenados na nuvem, $F = \{f_1, f_2, \dots, f_n\}, (1 \leq i \leq n)$.

- P é o conjunto de todos os diretórios p_i existentes na nuvem, $F = \{p_1, p_2, \dots, p_n\}, (1 \leq i \leq n)$.

- $R = \{r_1, r_2, \dots, r_n\}$ é o conjunto de todas as relações de posse que os usuários possuem sobre seus dados, tal que $r_i = \{ri_1, ri_2, \dots, ri_m\}$ $(1 \leq i \leq n)$ representa a relação de posse de cada arquivos e pasta pertencente a um usuário u_i da nuvem. O valor ri_k é a k-ésima relação de posse de u_i $(1 \leq k \leq m)$. Ele é definido pelo par $(u_i : f_i)$ e pelo par $(u_i : p_i), f_i \in F$, $p_i \in P$ e $u_i \in U$.

Acima é introduzida a definição de nuvem de armazenamento adotada neste artigo. Ela é apresentada com base em quatro elementos existentes na nuvem: um conjunto de usuários, um conjunto de arquivos, um conjunto de diretórios e um conjunto de relacionamentos de posse de dados. Essa notação pode ser representada como um grafo bipartido direcionado em que um conjunto de vértices representa os usuários do serviço e outro conjunto distinto de vértices representam os dados dos usuários. As arestas do grafo representam o relacionamento de posse de arquivos e de diretórios que os usuários possuem.

Esta definição destaca a visão de colaboração existente nas nuvens de armazenamento, pois apresenta um conjunto de conexões de compartilhamento de arquivos que ocorrem quando dois ou mais usuários estão conectados a um mesmo nó. Através dela é possível abstrair esses serviços a partir dos principais componentes envolvidos no mecanismo proposto.

Definição 2. (Caixas de Interesses) Seja $P = \{p_1, p_2, \dots, p_n\}$ o conjunto de diretórios existentes na nuvem *CSP* e seja *CI* o subconjunto $CI = \{c_1, c_2, \dots, c_m\}$ $CI \subseteq P$ de todas as caixas de interesses da nuvem. O elemento $p_i \in P$ $(1 \leq i \leq n)$ é uma caixa

de interesse $c_i \in CI$ $(1 \leq i \leq m)$ se possuir os metadados gerais e atributos Att_j como metadados específicos $(Att_j \geq 1)$.

A Definição 2 caracteriza as estruturas definidas neste artigo como caixas de interesses. Ela captura a noção de que essas caixas são criadas a partir de metadados gerais e metadados específicos definidos nelas. Esses metadados são importantes em dois aspectos. O primeiro é que eles distinguem as caixas de interesses de outras estruturas (ex. pastas). O segundo é que através deles o usuário é capaz de refinar o controle de acesso sob seus arquivos.

Conforme apresentado na Subseção 4.1, os atributos são disponibilizados pelos serviços de identidade federada. Considera-se então que cada usuário u_i apresenta um conjunto de atributos na federação. Esse conjunto é representado como um vetor Att_j que possui a forma: $[A_1(i), A_2(i), \ldots A_n(i)]$. $A_k(i)$ é o k-ésimo atributo de u_i $(1 \leq k \leq n)$. Cada atributo é representado por um par no formato *<propriedade: valor>*, $A_k(i) = < Prop_i = Pval_i >$. Nos casos em que um usuário possui mais de um valor para um mesmo atributo (atributo multivalorado), considera-se um par para cada valor, isto é, $< Prop_i = Pval_i > e < Prop_i = Pval_{i+1} >$ indicam que $Prop_i$ possui os valores $Pval_i$ e $Pval_{i+1}$.

Portanto, para fins de exemplificação, uma caixa de interesse c_i de um usuário u_i que possui o atributo conforme a Figura 2, possui o par $< eduPersonAffiliation = student >$ como metadado.

```
<saml:Attribute
      Name="urn:oid:1.5.7.1.4.1.5453.3.3.2.1"
      FriendlyName="eduPersonAffiliation">
      <saml:AttributeValue>
    Student</saml:AttributeValue>
</saml:Attribute>
```

Figura 2. Exemplo de Atributo de Federação de Identidade (Exemplo Simplificado).

Definição 3. (Compartilhamento Baseado em Interesses em Comum) Seja c_i e c_j duas caixas de interesses pertencentes a usuários distintos. Seja *CIFi e CIFj* o conjunto de arquivos existentes em c_i e c_j respectivamente. O compartilhamento baseado em interesses em comum ocorre quando *CIFi* ∩ *CIFj* ≠ ∅. Então, c_i será adicionada dos elementos *CIFj - CIFi* ∩ *CIFj*, e c_j será adicionada dos elementos *CIFi - CIFi* ∩ *CIFj*.

A definição acima apresenta a noção de compartilhamento baseado em interesses em comum introduzida neste artigo. Em outras palavras, esse tipo de compartilhamento ocorre quando os usuários envolvidos apresentam um ou mais arquivos em comum nas suas caixas de interesses. Satisfeita essa condição, os arquivos distintos em ambas podem ser compartilhados com o outro usuário.

A concepção por trás desta forma de compartilhamento é que os usuários que possuem arquivos iguais entre si, agrupam arquivos de um mesmo assunto (ou assuntos similares) em suas caixas de interesses. Portanto, os arquivos compartilhados podem se referi a um interesse em comum dos usuários, isto é, o interesse por um mesmo assunto.

4.3. Funcionamento do Mecanismo

Nesta seção são descritas as etapas introduzidas na Seção 4.1. Nelas são utilizadas notações para o detalhamento do mecanismo. Essas notações são apresentadas na tabela abaixo:

Tabela 1. Notações dos elementos envolvidos na descrição do mecanismo.

CSP	Serviço de armazenamento de dados em nuvem. Ele é o *SP* da federação.	cf_i	Arquivo incluído em uma caixa de interesse. Ele será fragmentado em k partes pelo mecanismo de deduplicação para ser armazenado em nuvem $(1 \leq i \leq n, cf_i \in CIF, \lvert CIF \rvert = n)$
FedIDP	Serviço de identidade federada. Ele disponibiliza o *IDP* da federação.	cf_{ij}	Parte de cf_i que será enviada para a nuvem, $0 \leq j \leq k$.
CIGer	Serviço gerenciador de caixas de interesses.	$H(cf_i)$	Valor *hash* de cf_i.
ID, Att	Credencial de acesso e atributos do usuário.	$H(cf_{ij})$	Valor *hash* de cf_{ij}.
Dedup(.)	Função de deduplicação de dados.	$CIcomp$	Conjunto de caixas de interesses que possuem os mesmos atributos que c_i, $CIcomp \subseteq CI$
$H(.)$	Função *Hash* aplicada sobre o arquivo e seus fragmentos. Ela pode ser compreendida como função que gera um *token* único que representa um arquivo.	$CIselec$	Conjunto de caixas de interesses selecionadas por u_i, $CIselec \subseteq CIcomp$.
u_i	Usuário do serviço de armazenamento de dados em nuvem $(1 \leq i \leq n,\ u_i \in U, \lvert U \rvert = n)$.	$Hcomp$	Conjunto de todos os $H(cf_{ij})$ de $CIselec$.
c_i	Caixa de Interesses de um usuário u_i $(1 \leq i \leq n,\ c_i \in CI, \lvert CI \rvert = n)$.	$Fcomp$	Conjunto de arquivos recebidos através do compartilhamento $Fcomp \subseteq F$.

I. Etapa de Autenticação e Obtenção de Atributos

A autenticação e obtenção de atributos consistem na primeira etapa do mecanismo. Nela o usuário autentica-se utilizando o serviço de identidade federada. Após isso, ele obtém seus atributos na federação assim como o acesso a nuvem. O procedimento de autenticação ocorre conforme o padrão adotado pelo protocolo SAML. A Figura 3 sintetiza esta etapa:

Operação 1: $u_i \rightarrow$ *CSP*
Operação 2: CSP $\xrightarrow{u_i}$ *FedIDP*
Operação 3: FedIDP $\xrightarrow{Att,\ ID}$ u_i
Operação 4: $u_i \xrightarrow{Att,\ ID}$ *CSP*

Figura 3. Etapa de Autenticação e Obtenção de Atributos (Resumo)

Inicialmente o usuário u_i solicita acesso ao serviço de armazenamento de dados em nuvem *CSP* (*Operação 1*). Então, a nuvem redireciona o usuário para um *IDP* da federação para que ele se autentique e obtenha acesso ao serviço (*Operação 2*). Obtendo sucesso na autenticação, o usuário recebe sua credencial e seus atributos (*Operação 3*) e então comunicasse com a nuvem novamente para acessar seu espaço de armazenamento (*Operação 4*).

II. Etapa de Criação de Caixas de Interesses

Após a Etapa 1, o usuário encontra-se autenticado e também possui seus atributos. Nesta segunda etapa ele se comunica com o serviço de armazenamento de dados em nuvem e define metadados em um diretório transformando-o em um caixa de interesse. A partir de então, quando o usuário incluir arquivos neste diretório, o serviço de nuvem utiliza o componente CIGer para registrar os arquivos pertencentes a ele.

A Figura 4 apresenta o funcionamento desta etapa.

Operação 1: $u_i \xrightarrow{c_i}$ *CSP*
Operação 2: CSP $\xrightarrow{u_i,\ c_i,\ att}$ *CIGer*
Operação 3: $u_i \xrightarrow{H(cf_i),\ H(cf_{ij})}$ *CSP*
Operação 4: CSP computa:
$Dedup(cf_{ij}), \forall j, 0 \leq j \leq k;$
Operação 5: CSP $\xrightarrow{H(cf_i),\ H(cf_{ij})}$ *CIGer*

Figura 4. Etapa de Criação de Caixas de Interesses (Resumo)

A etapa é iniciada pelo usuário quando ele cria uma caixa de interesses na nuvem (*Operação 1*). Para isso ele seleciona um diretório na nuvem e define seus metadados. Os metadados gerais que podem ser definidos são nome, data de criação e descrição para caixa de interesse. Os metadados específicos são os atributos Att da federação. Em seguida, C*SP* informa o componente *CIGer* que o usuário u_i criou a caixa de interesse c_i assim como os atributos definidos nela (*Operação 2*). *CIGer* registra esses dados para manter o conhecimento das caixas de interesses criadas.

Após definir uma caixa de interesse, o usuário pode seguir o procedimento de envio de arquivos de forma idêntica ao envio para um diretório qualquer. Devido ao mecanismo de deduplicação da nuvem, o arquivo é fragmentado em pedaços antes de ser enviado e também são calculados os valores *hash* $H(cf_i)$ e $H(cf_{ij})$. Esses valores são, então, enviados para a nuvem *CSP (Operação 3)*.

Ao receber $H(cf_i)$ e $H(cf_{ij})$, *CSP* verifica se esses dados já existem na nuvem aplicando a deduplicação de dados (*Operação 4*). O *CSP* termina esta etapa enviando $H(cf_i)$ e $H(cf_{ij})$ ao componente *CIGer*. Desta forma, este componente mantem o controle dos arquivos existentes nas caixas de interesses da nuvem registrando as relações de posse daquele usuário (isto é, o par $(u_i: f_i)$).

Caso o arquivo não seja deduplicado, o *CSP* apenas envia $H(cf_i)$ e $H(cf_{ij})$ ao componente *CIGer* e não realiza as etapas seguintes. Isso porque o arquivo ainda não existe na nuvem e portanto não existem caixas de interesses com arquivos em comum para realizar o compartilhamento. Caso contrário a Etapa 3 é iniciada.

III. Etapa de Identificação de Usuários

Na etapa de identificação de Usuários apenas o *CSP* e o componente *CIGer* estão envolvidos. De forma geral, ela consiste em um processamento realizado pelo *CIGer* para a identificação de caixas de interesses que possuem um arquivo em comum e na notificação da existência dessas caixas para a nuvem.

Se *CSP* encontrou armazenado todos os $H(cf_{ij}), \forall j, 0 \leq j \leq k$

Operação 1: CIGer procura pelas caixas de interesses associadas a $H(cf_i)$ que possuam atributos att iguais ao de c_i de u_i.
Operação 2: CIGer $\xrightarrow{CIcomp}$ *CSP*

Figura 5. Etapa de Identificação de Usuários (Resumo)

A identificação de potenciais usuário para o compartilhamento ocorre através da busca por caixas de interesses definidas por outros usuários da nuvem. Essa busca e realizada através do valor $H(cf_i)$ (*Operação 1*). Ao encontra *o hash* do arquivo, o componente *CIGer* obtém as caixas que também possuem o mesmo arquivo e atributos em comum.

Isso é possível, pois, conforme apresentado na descrição das etapas anteriores, o CIGer é notificado sobre a existência de uma caixa de interesses toda vez que ela é criada, assim como é notificado sobre arquivos que ela possui no momento em que eles são adicionados nela.

Tendo resgatado as caixas de interesses em comum, CIGer envia essa informação ao *CSP* e então se inicia a Etapa 4.

IV. Compartilhamento de Dados

Esta consiste na última etapa do mecanismo. Nela o componente CIGer notifica o serviço de armazenamento em nuvem quando encontrar caixas de interesses com arquivos em comum. A nuvem por sua vez, notifica o usuário e, mediante seu aceite, promove o compartilhamento de dados de uma caixa de interesse para outra.

A Figura 6 apresenta o protocolo de funcionamento desta etapa.

Operação 1: CSP $\xrightarrow{CIcomp}$ u_i
Operação 2: $u_i \xrightarrow{CIselec}$ *CSP*
Operação 3: CSP $\xrightarrow{CIselec}$ *CIGer*

Operação 4: CIGer $\xrightarrow{HComp}$ *CSP*

Operação 5:

CSP obtém para cada elemento de $HComp$

cf_{ij} a partir de cada $H(cf_{ij})$ $\forall i, 1 \leq i \leq n$

cf_i a partir de cada cf_{ij}, $\forall i, 1 \leq i \leq n$

Operação 6: CSP $\xrightarrow{Fcomp}$ u_i

Figura 6. Etapa de Compartilhamento de Dados (Resumo)

Inicialmente, o serviço de nuvem realiza a *Operação 1* do fluxo acima. Esta operação refere-se a disponibilização da lista de caixas de interesses encontradas na Etapa III para o usuário que adicionou um arquivo na sua caixa de interesses conforme apresentado na Etapa II. O serviço de nuvem *CSP* sugere as caixas encontradas e aguarda o aceite do usuário. O usuário por sua vez, pode aceitar apenas aquelas que desejar. Desta forma, apenas a solicitação de compartilhamento com as caixas de interesses aceitas por u_i são encaminhadas para a nuvem (*Operação 2*).

Recebendo as caixas de interesses que o usuário deseja ter disponíveis para si, o *CSP* realiza a *Operação 3* e as encaminha para CIGer. O componente CIGer procura as caixas de interesse contidas na lista recebida e retorna os valores *hashs* dos fragmentos referentes aos arquivos incluídos nelas (*Operação 4*). Na *Operação 5*, *CSP* toma conhecimento do(s) arquivo(s) cf_i que será(m) disponibilizado(s) ao usuário encontrando os fragmentos cf_{ij} a partir dos valores $H(cf_{ij})$. Então, por fim, C*SP* disponibiliza os arquivos pertencentes as caixas de interesses aceitadas por u_i *(Operação 6)*. Antes do envio dos arquivos, a nuvem se sincroniza com o usuário (ex. comunicando-se com um aplicativo cliente) e verifica quais arquivos já existem na caixa de interesse dele, enviando desta forma apenas aqueles que não existem nela.

5. DISCUSSÃO INICIAL

Conforme se observa na Seção 4, o mecanismo proposto apresenta alguns elementos essenciais para sua realização. A nuvem de armazenamento e a federação de identidade, em particular, possuem papel fundamental nesse contexto, pois atuam como o repositório de execução do compartilhamento e como o centro provedor de atributos respectivamente. Desta forma, algumas considerações relativas ao funcionamento e ao cenário de implantação do mecanismo por meio desses elementos são apresentadas nesta seção.

A primeira consideração está relacionada com a forma de deduplicação utilizada. De acordo com o apresentado na Seção 3, a deduplicação pode ocorrer de duas formas. No compartilhamento por caixas de interesses considera-se a deduplicação em nível de blocos. Essa decisão está em concordância com o tipo de deduplicação que vem sendo utilizada pelos serviços de armazenamento nuvem, como é o caso, por exemplo, do Dropbox [8], e do Mozy [23]. Assim, o mecanismo proposto possui aderência ao contexto dos serviços de nuvem possibilitando maior facilidade de implantação em um cenário real, como por exemplo, através em uma plataforma de nuvem de código aberto.

Além disso, o mecanismo pode ser facilmente adaptado para ser executado com a abordagem de deduplicação em nível de arquivo. Para isso, é necessário que o mecanismo inicie a Etapa III caso encontre diretamente $H(cf_i)$. e não os valores $H(cf_{ij})$. Desta forma há uma maior flexibilidade na aplicação do compartilhamento em outros cenários, como por exemplo, outros serviços de armazenamento de dados que utilização a deduplicação de arquivos inteiros.

Outra consideração está relacionada com a utilização de federações de identidade como centro de distribuição de atributos. Esses serviços têm sido amplamente adotados por diversas instituições [29][30][31] e são capazes de oferecer tanto a autenticação quanto os atributos através de um único serviço diminuindo a complexidade gerenciamento de ambos.

Além disso, a utilização desses serviços permite que a implantação do mecanismo seja flexibilizada para nuvens com diferentes modelos de implantação. Em outras palavras, através da utilização de federações de identidade o mecanismo de compartilhamento por caixas de interesses pode ser implantado tanto no contexto de nuvens privadas e comunitárias quanto no de nuvens públicas.

Em nuvens privadas e comunitárias, é desejável que apenas os usuários das organizações envolvidas possam realizar o compartilhamento. Neste cenário pode se utilizar um serviço de identidade federada mantido pelas organizações, limitando o escopo de compartilhamento entre seus usuários. Um possível cenário de aplicação nesse contexto seria a implantação do compartilhamento por caixas de interesses em uma nuvem comunitária de pesquisa. Os seus usuários seriam professores e pesquisadores que armazenam suas pesquisas (ex. artigos publicados) nas suas caixas de interesses e assim colaborariam compartilhando-as com outros pesquisadores que pesquisam sobre um tema em comum.

Em contra partida, as federações organizacionais não contemplam o cenário de utilização de nuvens públicas em que há a necessidade de compartilhamento entre usuários do público em geral. Neste caso é necessário um serviço de identidade federada público como centro provedor de atributos. Para atender esse contexto, o mecanismo pode ser implantado para atuar utilizando as redes sociais como centro provedor de atributos. Isso é possível pois grande parte delas também utilizam protocolos de autenticação federada, como o Oauth [32], para fornecer a autorização de acesso a recursos de seus usuários para terceiros. Exemplificando com a rede social Facebook[5], os recursos disponibilizados são as propriedades dos perfis de usuários [33], tais quais sua localização e o seu local de estudo, por exemplo. Essas propriedades seriam os atributos que os usuários adicionariam as suas caixas de interesses. Através desses novos centros de distribuição de atributos o mecanismo amplia sua abrangência viabilizando o compartilhamento por interesses em comum além do contexto organizacional.

6. CONCLUSÃO

Neste artigo foi proposto um novo mecanismo para o compartilhamento de dados entre usuários desconhecidos de serviços de armazenamento de dados em nuvem. O compartilhamento por caixas de interesses usufrui da técnica de deduplicação utilizada pelo provedor de nuvem e fornece a capacidade de compartilhamento controlado de arquivos. Desta forma, os provedores não só se beneficiam da utilização racional de recursos de armazenamento, como também podem oferecer uma forma complementar de colaboração aos seus usuários.

Por fim, ressalta-se que este trabalho é apenas o esforço inicial de se obter um mecanismo que viabilize o compartilhamento de dados de forma mais ampla entre os usuários. Como trabalhos futuros, prevê-se o desenvolvimento de um protótipo do mecanismo proposto. A partir dele, então, a realização de um estudo sobre diversos aspectos, tais quais, a avaliação da eficácia de colaboração, através do teste do mecanismo em um cenário real com um grupo de usuários, a análise de desempenho do mecanismo e a capacidade de estender o funcionamento do mecanismo utilizando serviços de nuvem diferentes, como por exemplo, nuvens híbridas. Por meio deste estudo poderão ser identificadas possíveis melhorias e extensões para a proposta.

7. REFERÊNCIAS

[1] Borgmann M., Hahn T., Herfert M., Kunz T., Richter M., Viebeg U. e Vowe S., "On the Security of Cloud Storage Services," 2012. [Online]. Available: https://www.sit.fraunhofer.de/fileadmin/dokumente/studienund_technical_reports/Cloud-Storage-Security_a4.pdf. [Acesso em abril 2014].

[2] Mell. P. e Grance. T., "The NIST Definition of Cloud Computing," 2011. [Online]. Available: http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf.

[3] "Overview of Managing Access - Amazon S3," [Online]. Available: http://docs.aws.amazon.com/AmazonS3/latest/dev/access-control-overview.html. [Acesso em abril 2014].

[4] "Role Based Access Control - Rackspace Cloud Files," [Online]. Available: http://docs.rackspace.com/files/api/v1/cf-devguide/content/RBAC.html. [Acesso em abril 2014].

[5] "Facebook," [Online]. Available: http://www.facebook.com. [Acesso em agosto 2014].

[6] "ResearchGate," [Online]. Available: http://www.researchgate.net/. [Acesso em agosto 2014].

[7] "Academia.edu," [Online]. Available: http://academia.edu. [Acesso em agosto 2014].

[8] "Dropbox," [Online]. Available: http://www.dropbox.com. [Acesso em abril 2014].

[9] "Box," [Online]. Available: http://www.box.com. [Acesso em abril 2014].

[10] "Google Drive," [Online]. Available: drive.google.com. [Acesso em abril 2014].

[11] Diniz, T. F.S., Silva, C. E. e Araujo R., "Integrando o Openstack Keystone com Federações de Identidade," *Simpósio Brasileiro em Segurança da Informação e Sistemas Computacionais,* pp. 465-474 , 2013.

[12] Silva L., Silva F., Araujo R., Diniz, T. F.S. e Silva, C. E., "Estudo de Caso: Integração de Clientes de Nuvem Openstack Swift com Federação de Identidade," *Simpósio Brasileiro em Segurança da Informação e Sistemas Computacionais,* pp. 455-464 , 2013.

[13] Kolouzis S., Cuching R., Bellloum A. e Bubak, M., "Cloud Federation for Sharinh Scientific Data,", *8th* International eScience 2012 Conference on, 2012.

[14] K. Chard, K. Bubendorfer, S. Caton e O. F. Rana, "Social Cloud Computing: A Vision for Socially Motivated Resource Sharing," *Services Computing, IEEE Transactions on,* pp. 551 - 563, 2012.

[15] Thaufeeg A. M., Bubendorfer K. e Chard K., "Collaborative eResearch in a Social Cloud," *E-Science (e-Science), 2011 IEEE 7th International Conference on,* pp. 224 - 231, 2011.

[16] John K., Bubendorfer K. e Chard K., "A Social Cloud for Public eResearch," *E-Science (e-Science), 2011 IEEE 7th International Conference on,* pp. 363 - 370, 2011.

[17] Punceva M., Rodero I., Parashar M., Rana O. F. e Petri I., "Incentivising Resource Sharing in Social Clouds," *Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), 2012 IEEE 21st International Workshop on,* pp. 185 - 190, 2012.

[18] Christin D., López P. S., Reinhardt A. e Hollick M., "Share with strangers: Privacy bubbles as user centered privacy control for mobile content sharing applications," *Information Security Technical Report,* p. 105–116, 2013.

[19] C. Bo, Z. F. Li e W. Can, "Research on Chunking Algorithms of Data De-duplication," *Proceedings of the 2012 International Conference on Communication, Electronics and Automation Engineering,* pp. 1019-1025, 2013.

[20] D. Mishra e S. Sharma, "Comprehensive study of data de-duplication," *International Conference on Cloud, Big Data and Trust 2013,* 2013.

[21] D. Meister, J. Kaiser, A. Brinkmann, T. Cortes, M. Kuhn e J. Kunkel, "A study on data deduplication in HPC storage systems," *High Performance Computing, Networking, Storage and Analysis (SC), 2012 International Conference for,* pp. 1-11, 2012.

[22] K. Jin e E. L. Miller, "The Effectiveness of Deduplication on Virtual Machine Disk Images," *SYSTOR '09 Proceedings of SYSTOR 2009: The Israeli Experimental Systems Conference,* 2009.

[23] "Mozy" [Online]. Available: https://mozy.com/. [Acesso em abril 2014].

[24] "Memopal" [Online]. Avaiable: http://www.memopal.com/pt-br/. [Acesso em abril 2014]

[25] "Secure Hash Standart". [Online] Available: http://csrc.nist.gov/publications/fips/fips180-4/fips-180-4.pdf [Acesso em abril 2014]

[26] "SAML Especifications," [Online]. Available: http://saml.xml.org/saml-specifications. [Acesso em abril 2014].

[27] "CAS Federation," [Online]. Available: http://www.uky.edu/ukit/iog/cas. [Acesso em abril 2014].

[28] "Shibbolleth," [Online]. Available: https://shibboleth.net/. [Acesso em abril 2014].

[29] “CAFe - Comunidade Acadêmica Federada” [Online]. Available: https://portal.rnp.br/web/servicos/cafe. [Acesso em abril 2014].

[30] “CANARIE - Canadian Access Federation,” [Online]. Available: http://www.canarie.ca/en/caf/join. [Acesso em abril 2014].

[31] “InCommon Federation,” [Online]. Available: http://www.incommonfederation.org/. [Acesso em abril 2014].

[32] Hardt D. “The Oauth Autorization Framework”. Available: http://tools.ietf.org/html/rfc6749 [Acesso em abril 2014].

[33] “Facebook Graph API Reference User”. [Online] Available: https://developers.facebook.com/docs/graph-api/reference/v2.0/user[Acesso em abril 2014]