UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FELIPE LEITE DA SILVA

CAIXAS DE INTERESSES: UM NOVO MECANISMO PARA A COLABORAÇÃO ATRAVÉS DE NUVENS DE ARMAZENAMENTO DE DADOS

Belém-PA
2015

FELIPE LEITE DA SILVA

CAIXAS DE INTERESSES: UM NOVO MECANISMO PARA A COLABORAÇÃO ATRAVÉS DE NUVENS DE ARMAZENAMENTO DE DADOS

Dissertação de Mestrado apresentada para a obtenção do grau de mestre em Ciência da Computação no Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas e Naturais da Universidade Federal do Pará.
Área de Concentração Redes de Computadores.
Orientador Prof. Dr. Nelson Cruz Sampaio Neto

Belém-PA
2015

FELIPE LEITE DA SILVA

CAIXAS DE INTERESSES: UM NOVO MECANISMO PARA A COLABORAÇÃO ATRAVÉS DE NUVENS DE ARMAZENAMENTO DE DADOS

Dissertação de Mestrado apresentada para a obtenção do grau de mestre em Ciência da Computação no Programa de Pós Graduação em Ciência da Computação do Instituto de Ciências Exatas e Naturais da Universidade Federal do Pará.

Aprovado em ___/____/____

BANCA EXAMINADORA

_____

Prof. Dr. Nelson Cruz Sampaio Neto
Faculdade de Computação - Instituto de Ciências Exatas e Naturais- UFPA – Orientador

_____

Prof. Dr. Gláucio Haroldo Silva de Carvalho
Faculdade de Computação - Instituto de Ciências Exatas e Naturais- UFPA – Membro

_____

Prof. Dr. Josivaldo de Souza Araújo
Programa de Pós Graduação em Ciência da Computação - UFPA – Membro Interno

_____

Prof. Dr. Roberto Samarone dos Santos Araújo
Faculdade de Computação - Instituto de Ciências Exatas e Naturais- UFPA – Membro Externo

AGRADECIMENTOS

Este trabalho não poderia ser concluído sem a atuação grandiosa de Deus na minha vida e a cooperação de diversas pessoas que estiveram presente nos momentos de dificuldade e noites de sono mal dormidas.

Primeiramente, agradeço a Deus por sempre me fortalecer e por ter mantido acesa a confiança de que essa etapa da minha vida seria concluída com êxito. Graças a Ele mesmo nos dias de maiores dificuldades houve um dia seguinte esperando para me entregar uma nova dose de ânimo e paciência.

À minha família: Margareth Leite de Melo e Thiago Leite da Silva, minha mãe e meu irmão respectivamente, por sempre me oferecer uma vida tranquila enquanto atravesso as turbulências do dia a dia.

Ao prof. Nelson Cruz Sampaio Neto que me orientou neste trabalho e ao prof. Roberto Samarone dos Santos Araújo, que também me auxiliou em muitos momentos de dificuldades. Obrigado por sempre estarem dispostos a me ensinar. Obrigado pela confiança e paciência.

À Amália Honda que suportou manhãs, tardes e noites de reclamações oriundas de dificuldades na pesquisa. Em contrapartida me retornou com compressão, incentivo e companheirismo.

Aos professores do programa de pós-graduação em ciência da computação da Universidade Federal do Pará. Todos contribuíram ao longo de cada semestre para que esse trabalho obtivesse sucesso.

A todos aqueles que mesmo não tendo citado o nome foram importantes para o êxito deste trabalho.

RESUMO

Os serviços baseados em computação em nuvem vêm sendo adotados cada vez mais por indivíduos e organizações que buscam obter recursos computacionais de forma simples e sob demanda. Os serviços de armazenamento de dados em nuvem, em particular, tornaram-se uma das principais tendências nesse contexto. Eles oferecem diversos benefícios aos usuários, como a alta disponibilidade e o acesso multidispositivo aos seus arquivos. Além desses benefícios, as nuvens de armazenamento de dados caracterizam-se por facilitar a colaboração entre seus usuários. O compartilhamento de dados é um dos principais recursos ofertados para promover tal colaboração.

Embora o compartilhamento de dados em nuvens de armazenamento apresente benefícios, eles também limitam a colaboração entre os usuários, pois restringem o compartilhamento entre aqueles que já conhecem uns aos outros. Neste contexto, esta pesquisa apresenta uma proposta que objetiva mitigar esta limitação. A proposta, denominada de compartilhamento por caixas de interesses, consiste em um mecanismo que permite quaisquer usuários de um serviço de armazenamento de dados em nuvem compartilharem arquivos entre si desde que apresente potenciais interesse comum.

Nesta pesquisa são apresentados dois cenários de aplicação assim como um protótipo do mecanismo proposto. Através deste protótipo, foram realizados testes de desempenho visando verificar quais aspectos do mecanismo interferem no tempo de sua execução.

PALAVRAS-CHAVE: Computação em Nuvem, Armazenamento de dados em Nuvem, Compartilhamento de Dados, Colaboração

LISTA DE ILUSTRAÇÕES

LISTA DE GRÁFICOS

LISTA DE QUADROS

LISTA DE SIGLAS

ABREVIAÇÃO	*TERMO*
ACM	*Association for Computer Machine*
ACL	*Access Control List*
CAFÉ	Comunidade Acadêmica Federada
CNO	*Collaborative Networks of Organizations*
FIM	*Federated Indentity Management*
IAAS	*Infrastructure as a Service*
IBAC	*Identity Based Access Control*
IDP	*Identity Provider*
NREN	*National Research and Educational Networks*
OASIS	*Organization for the Advancement of Structured Information Standards*
PAAS	*Platform as a Service*
RBAC	*Role Based Access Control*
RNP	Rede Nacional de Pesquisa
SAAS	*Software as a Service*
SAML	*Security Assertion Markup Language*
SP	*Service Provider*
SSO	*Single Sign On*
URL	*Uniform Resource Locator*
WBSN	*Web Based Social Network*
XML	*eXtensible Markup Language*

SUMÁRIO

1. INTRODUÇÃO

A quantidade de dados digitais tem crescido rapidamente com o surgimento de novas tecnologias e o aprimoramento dos dispositivos móveis. Indivíduos equipados com diferentes aparelhos produzem, por exemplo, grandes quantidades de mídias, como fotos e vídeos. As organizações, por sua vez, produzem diferentes arquivos para desenvolver suas atividades. Esse cenário motivou o surgimento de novos serviços que atendessem à demandas cada vez maiores por armazenamento de dados.

As nuvens de armazenamento surgem nesse contexto como uma proposta para solucionar tal demanda. Elas se caracterizam por fornecer uma interface amigável e de fácil acesso a ambientes virtuais remotos de armazenamento de dados. Essas nuvens oferecem um serviço de rápido provisionamento, custos baseados nas necessidades dos seus clientes e incluem a alta disponibilidade (ex. arquivos são acessíveis a qualquer momento) e o acesso multiplataforma por meio de diferentes dispositivos (BORGMANN, 2012) (MELL e GRANCE, 2011).

Dentre os seus diversos benefícios, os serviços de armazenamento de dados em nuvem caracterizam-se também por oferecer diferentes recursos que facilitem a colaboração entre seus usuários. O compartilhamento de dados, em particular, é amplamente adotado para promover tal colaboração (EUROSTATS, 2014). Por meio dele, os usuários podem trocar grandes quantidades de dados entre si e trabalhar de forma integrada. Por exemplo, organizações podem compartilhar dados dinamicamente entre suas filiais e pesquisadores podem cooperar em trabalhos multi-institucionais.

Tradicionalmente, as nuvens de armazenamento realizam os mecanismos de compartilhamento através do modelo de controle de acesso baseado em identidade (*Identity Based Access Control - IBAC*) (KHAN, 2012) (MAJUMDER, NAMASUDRA e NATH, 2014). Em outras palavras, nesses mecanismos o usuário deve determinar 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[1], e os níveis hierárquicos (*Role Based Access Control - RBAC*) como utilizados pela nuvem Rackspace[2] (AUSANKA-CRUES, 2004). O compartilhamento baseado em tais modelos de controle de acesso 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 limitações inerente quando utilizados juntamente com nuvens de armazenamento (YOUNIS, KIFAYAT e MERABTI, 2014). No contexto do compartilhamento de dados, a necessidade de conhecimento prévio do identificador do usuário receptor do compartilhamento configura-se como uma limitação. Eles viabilizam apenas 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 contemplados por estes mecanismos.

Por outro lado, as nuvens de armazenamento geralmente possibilitam a criação de arquivos ou pastas públicas como alternativa aos modelos de controle de acesso baseado em identidade. Essas formas de compartilhamento são observadas em serviços como Dropbox (2015) e Googel Drive (2015), por exemplo. Nesse tipo de compartilhamento, denominado de publicação de arquivo, o serviço de nuvem gera uma *URL* que representa o arquivo do usuário. Essa *URL* é disponibilizada ao usuário portador do arquivo que, por sua vez, a concede a outra (s) pessoa (s).

Tal solução amplia a possibilidade de compartilhamento de arquivos dos usuários da nuvem. Isso porque ela se baseia na remoção de restrições de acesso sobre o item selecionado tornando-o disponível ao público em geral. No entanto, a publicação de arquivos compromete o controle do usuário sobre seus dados, uma vez que limita a forma de definir restrições de acesso ao arquivo.

Uma solução comumente adotada para mitigar essa limitação é a disponibilização do arquivo em plataformas externas (ex. sites pessoais, redes sociais). Os serviços de

[1] https://aws.amazon.com/s3/
[2] https://www.rackspace.com/

armazenamento em nuvem disponibilizam o arquivo através de uma *URL* pública e o usuário, por sua vez, a disponibiliza a outros indivíduos publicando-a na plataforma desejada.

Entretanto, essa abordagem aumenta a complexidade de gerenciamento dos arquivos disponibilizados. Isso porque o usuário deve gerenciar cada *URL* publicada na correspondente plataforma em que foi inserida. Outro problema está na dificuldade de promover o compartilhamento com indivíduos que não possuem relacionamento com o portador do arquivo. Pessoas que vivem em localidades com acesso restrito a sites de pesquisa e a redes sociais, por exemplo, dificilmente encontrariam o link publicado.

Mediante esse cenário, este trabalho propõe uma nova forma de colaboração entre usuários de serviços de armazenamento de dados em nuvem. Ele introduz o mecanismo de *compartilhamento por caixas de interesses*.

O mecanismo proposto facilita o compartilhamento de arquivos entre quaisquer usuários dentro de um serviço de armazenamento em nuvem. Para tal, ele é capaz de identificar potencias usuários com interesse em comum e viabilizar que eles troquem arquivos entre si. Adicionalmente, o usuário é capaz de gerenciar o acesso sobre seus dados compartilhados. Ele pode restringi-lo utilizando atributos pessoais providos por um centro provedor de atributos. Desta forma, este mecanismo amplia o escopo de compartilhamento de dados em nuvem sem comprometer o controle do usuário sobre seus dados.

1.1 Motivação

A computação em nuvem tem ganhado força nos últimos anos. Pesquisas recentes apresentam esse modelo computacional como uma das principais tendenciais tecnológicas dentro das estratégias organizacionais (GARTNER, 2015) (COLUMBUS, 2015)

O armazenamento de dados destaca-se dentre os diversos tipos de serviços aderentes ao modelo de nuvem. Esse tipo de serviço vem se tornando cada vez mais popular na medida em que há um aumento na quantidade de arquivos produzidos por todos os

setores da sociedade.

As nuvens de armazenamento ganham visibilidade devido aos diversos benefícios que oferecem aos seus usuários. O acesso multiplataforma, a capacidade de compartilhar arquivos e interagir colaborativamente com outros usuários e a capacidade de gerenciar dados armazenados a qualquer momento, são exemplos desses benefícios.

O compartilhamento de dados, em particular, é um dos recursos mais atraentes dos serviços de armazenamento em nuvem (EUROSTATS, 2014). Por meio dele, é possível que os usuários interajam e colaborem uns com os outros de diversas formas, por exemplo, trocando fotos com familiares, trabalhando de forma integrada com uma equipe em outra localização geográfica e disponibilizando arquivos de forma dinâmica para grupos de usuários específicos.

Embora os mecanismos de compartilhamento de dados tradicionais possuam seus benefícios, o crescente aumento na produção de dados levou o surgimento de novos desafios no contexto de como os usuários buscam e oferecem dados de seu interesse (GOLLU, SAROIU e WOLMAN, 2007). As redes sociais como o Facebook e Google+, por exemplo, são tecnologias que surgiram motivadas pelo interesse dos usuários de compartilhar dados pessoais com as pessoas com que se relacionam.

No contexto das nuvens de armazenamento também surgem desafios em relação a como os usuários compartilham seus dados. A confidencialidade no compartilhamento, o controle de acesso sobre os dados compartilhados e a busca por estratégias de compartilhamento que aprimorem a colaboração entre os usuários de serviços em nuvem são exemplos desses desafios.

Este último desafio, especificamente, está relacionado com a forma que os mecanismos atuais funcionam. Os mecanismos de compartilhamento tradicionais utilizam o controle de acesso baseado em identidade. Tal modelo dificulta a delegação de acesso aos dados em serviços com grandes quantidades de usuários que possuem perfis heterogêneos (KHAN, 2012). O compartilhamento de dados baseado nesse modelo, por sua vez, limita a troca de informações entre conjuntos de usuários que possuem identificadores uns dos outros.

Tendo em vista a carência de pesquisas ou de soluções que tratem desse desafio, este trabalho tem o objetivo de propor um mecanismo que possibilite o usuário usufrui do compartilhamento de dados com quaisquer outros usuários da nuvem independente

de possuir algum relacionamento com essa pessoa. Adicionalmente, a proposta visa um mecanismo em que o controle sobre os dados compartilhados está centrado no usuário de tal forma que ele possa definir restrições de acesso sobre seus arquivos.

1.2 Objetivos

Este trabalho tem como objetivo geral propor um mecanismo de compartilhamento controlado de dados em nuvens de armazenamento que viabilize a troca de arquivos entre quaisquer usuários que possuam potenciais interesses em comum.

Para atender ao objetivo geral, os seguintes objetivos específicos são contemplados:

1. Realizar um levantamento bibliográfico relacionado ao tema abordado;
2. Investigar os principais mecanismos relativos ao compartilhamento de dados em nuvem;
3. Elaborar um mecanismo que possibilite o compartilhamento controlado de dados entre usuários desconhecidos que possuem potenciais interesse em comum;
4. Desenvolver um protótipo do mecanismo como prova de conceito;
5. Apresentar considerações sobre aspectos de desempenho do mecanismo;
6. Descrever diferentes cenários de aplicação do mecanismo proposto no contexto de nuvens de armazenamento.

1.3 Metodologia de Pesquisa

Os autores Silva e Menezes (2001) descrevem a metodologia da pesquisa científica como um conjunto de etapas ordenadamente dispostas que devem ser realizadas durante o processo de investigação da pesquisa. Tomando como base esse conceito, esta seção

expõe, inicialmente, o tipo da pesquisa realizada. Em seguida são descritos os passos realizados para a sua formulação e desenvolvimento.

1.4.1 Tipo de Pesquisa

Silva e Menezes (2001) definem diferentes formas de classificação da pesquisa científica. Especificamente essas classificações são: quanto à natureza, quanto à abordagem do problema, quanto aos objetivos e quanto aos procedimentos técnicos. Adicionalmente, Waine (2007) observa que "a pesquisa em Ciência da Computação (...) envolve na maioria dos casos a construção de um programa, de um modelo, de um algoritmo ou de um sistema novo" e nesse contexto apresenta classificações e métodos específicos para essa área.

Mediante as classificações apresentadas pelos referidos autores, este trabalho caracteriza-se como:

Quanto à natureza: *pesquisa aplicada*, pois o objetivo deste trabalho é solucionar um problema específico no contexto da computação em nuvem. Além disso, ele produz conhecimento para uma aplicação prática da pesquisa.

Quanto à abordagem do problema: *pesquisa quantitativa*, pois apresenta dados que podem ser quantificados e analisados.

Quanto aos objetivos: *pesquisa descritiva*, pois propõe um mecanismo cujas características, componentes e seus relacionamentos são descritos. Além disso, ela também pode ser considerada uma *pesquisa explicativa*, porque possui analises e discussões de comportamento; e testes do mecanismo proposto.

Quanto aos procedimentos técnicos: *pesquisa experimental com uso de dados sintéticos*, pois considera a criação e a simulação de um protótipo.

1.4.2 Etapas da Pesquisa

A realização deste trabalho ocorreu na forma de quatro etapas. Essas etapas são apresentadas a seguir.

A primeira etapa consistiu em um levantamento bibliográfico inicial. A coleta de dados foi realizada através de levantamento de trabalhos nos seguintes repositórios: *Association for Computer Machine* (ACM)[3], IEEE[4] e Elsevier[5]. Os artigos verificados possuíam data base mínima no ano de 2010. Tal data foi utilizada, pois a partir deste ano há um crescimento acelerado do modelo de nuvens computacionais (MOHAMED, 2012). Através deste levantamento foram identificados trabalhos relacionados inseridos no contexto de compartilhamento de dados em nuvem. No fim desta etapa, foram selecionados aqueles que possuíam aplicabilidade referente à abrangência do compartilhamento de dados.

Na segunda etapa, foi elaborado o mecanismo proposto neste trabalho. Nesta etapa realizou-se um estudo mais aprofundado sobre o mecanismo de deduplicação e o modelo de computação em nuvem, especificamente, as nuvens de armazenamento. Baseado nesta contextualização, elaborou-se o mecanismo em si descrevendo seus componentes e fluxos de operação.

Na terceira etapa do projeto foi desenvolvido o protótipo como prova de conceito do mecanismo proposto na etapa anterior. Através desse protótipo, foram realizados testes e análises sobre esse mecanismo.

Por fim, na quarta etapa, registraram-se todos os dados levantados e elaborados na forma escrita de dissertação.

[3] https://www.acm.org/
[4] https://www.ieee.org/
[5] http://www.elsevier.com/

1.4 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, como por exemplo, o Memopal (2015), o Dropbox (2015), e o Mozy (2015). Em alguns casos o compartilhamento é aprimorado viabilizando a edição simultânea e em tempo real de alguns formatos de arquivos, como são os casos dos serviços Box (2015) e Google Drive (2015).

No contexto científico, alguns autores apresentam propostas relacionadas à utilização colaborativa das nuvens de armazenamento. No Brasil, o grupo de trabalho Computação em Nuvem para a Ciência (GT-CNC) realiza trabalhos nesse contexto (DINIZ *et al.*, 2013) (SILVA *et al.*, 2013). Eles propõem o desenvolvimento de uma nuvem de armazenamento cientifica em que os usuários das instituições pertencentes à 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.* (2014) também considera a criação de federações de nuvens para promover o compartilhamento de dados científicos no contexto do projeto VPH-Share. Essa proposta tem por objetivo principal unificar recursos computacionais de diferentes infraestruturas e permitir que os usuários da federação armazenem dados médicos e colaborem entre si.

Observa-se, contudo, que nessas propostas o compartilhamento de dados possui uma característica em comum. Eles facilitam apenas a troca de arquivos entre conjuntos específicos de usuários.

Especificamente, no cenário comercial o compartilhamento é possível apenas se o usuário obtiver algum identificador do receptor do compartilhamento. Nos trabalhos baseados em federações, a limitação é a mesma. Um usuário compartilhará seus dados com outro apenas se possuir algo que identifique o receptor na nuvem. Isso ocorre, pois, as propostas baseiam-se no modelo de controle de acesso baseado em identidade, conforme introduzido na Seção 1.1.

Outros trabalhos destinam-se a estender a capacidade de colaboração em nuvem a um escopo mais abrangente de usuários. Alguns pesquisadores apresentam o conceito de nuvem social nesse contexto (THAUFEEG *et al.*, 2011) e (KOSHY, BUBENDORFER e CHARD, 2011), (CATON *et al.*, 2012), (CHARD *et al.*, 2012) e (PUNCEVA *et al.*, 2012).

Uma nuvem social é um modelo em quem recursos computacionais são compartilhados com base no relacionamento que os indivíduos possuem em uma rede social. Por exemplo, um usuário pode decidir se deseja compartilhar recursos com pessoas mais próximas (ex. familiares e amigos) ou se deseja compartilhar seus recursos com pessoas mais distantes (ex. amigos de amigos).

O principal benefício desse modelo está em viabilizar que indivíduos que não possuem relacionamento próximo possam realizar o compartilhamento entre si (ex. o usuário poderá compartilhar recursos com um amigo de um amigo mesmo se não o conhecer). Desta forma, através das nuvens sociais é possível ampliar o escopo de compartilhamento entre os usuários. Adicionalmente, o controle de acesso no compartilhamento é mantido, pois o usuário define com quais pessoas de sua cadeia de relacionamento ele deseja compartilhar seus recursos.

Apesar dos benefícios, o modelo de nuvens sociais foi elaborado com o objetivo de promover o compartilhamento de recursos computacionais básicos, como armazenamento e o processamento, entre usuários de uma rede social. Nenhum dos trabalhos prevê a utilização deste modelo para facilitar o compartilhamento de dados em nuvens de armazenamento.

Além disso, o modelo de nuvem social tem seu funcionamento baseado na confiança que o usuário possui nas pessoas com que se relaciona (PUNCEVA *et al.*, 2012). Segundo Caton *et al.* (2012), a confiaça no contexto das nuvens sociais é baseada em experiencias passadas e futuras com outro indivíduo. Tendo isso em vista, na medida em que o compartilhador se encontrar mais distante do candidato receptor dos recursos (ex. amigo de um amigo de um amigo), menor poderá ser a sua confiança para promover o compartilhamento. Assim, mesmo que o modelo possa ser aplicado no contexto do compartilhamento de dados em nuvem, ele não facilitaria a troca de arquivos entre usuários que não possuem laços sociais.

Christin *et. al.* (2013) apresentam a proposta de bolha privada para aprimorar a

capacidade de compartilhamento de dados entre usuários de dispositivos móveis. Essa proposta permite que pessoas troquem dados entre si sem que eles se conheçam. Para tal, basta que o usuário defina limites espaciais e temporais em que deseje compartilhar seus arquivos. Outros usuários podem acessar esses dados desde que estejam dentro dos limites definidos pelo compartilhador.

Apesar de não prever o uso de nuvens de armazenamento em seu trabalho, a proposta pode ser estendida para inclui sua utilização. Isso é possível, pois o mecanismo é realizado através de uma aplicação para dispositivos móveis. Desta forma, é necessário apenas que a aplicação disponibilize os arquivos armazenados na nuvem.

Considerando o contexto de nuvens, o principal benefício do mecanismo está na abrangência do compartilhamento. Por meio da bolha privada os usuários da nuvem poderiam facilmente compartilhar seus dados com outros ao seu redor, mesmo sem os conhecer.

Contudo, o recurso de controle de acesso apresentado por Christin *et al.* (2013) é limitado ao compartilhamento através do uso de dispositivos móveis. Tal limitação ocorre devido aos parâmetros espaço-temporais necessários para o compartilhamento. Esses parâmetros são utilizados para regular o nível de privacidade que o usuário deseja ao compartilhar seus arquivos e são obtidos com base na sua posição geográfica. A utilização de terminais fixos (ex. computadores *Desktops*) limitaria o usuário a abrir mão de sua privacidade para compartilhar dados com pessoas distantes geograficamente.

Até onde se pôde pesquisar, este trabalho é o primeiro a considerar um mecanismo que facilite o compartilhamento controlado de dados entre quaisquer usuários de um serviço de armazenamento nuvem. Em particular, ele introduz um mecanismo que permite tal compartilhamento baseando-se em potenciais interesses que os usuários possuem em comum.

1.5 Organização do Trabalho

Este documento está dividido em mais quatro capítulos que fornecem o

embasamento teórico sobre o trabalho, os detalhes de sua elaboração, as considerações sobre o mecanismo proposto e as conclusões obtidas. Os capítulos são:

- **Capítulo 2 – Contextualização e Terminologias do Trabalho**

  Este capítulo apresenta os principais conceitos relacionados a esta pesquisa. Nele são apresentados os conceitos de nuvens computacionais, armazenamento de dados em nuvem, deduplicação de dados e federações de identidade.

- **Capítulo 3 – O Novo Mecanismo de Compartilhamento de Dados**

  Este capítulo apresentada a proposta da pesquisa. Ela, que consiste em um mecanismo de compartilhamento de dados em nuvens computacionais, é apresentada em detalhes juntamente com o protótipo desenvolvido como prova de conceito.

- **Capítulo 4 – Considerações sobre o Mecanismo**

  Este capítulo apresenta as discussões sobre o mecanismo proposto. Nela são apresentadas adaptações, cenários de aplicação e resultados de testes realizados.

- **Capítulo 5 – Considerações Finais**

  Este capítulo apresenta as considerações finais sobre a pesquisa desenvolvida, assim como sugestões para trabalhos futuros.

2. CONTEXTUALIZAÇÃO E TERMINOLOGIAS DO TRABALHO

A proposta deste trabalho, conforme apresentado no Capítulo 1, refere-se a um mecanismo destinado ao compartilhamento em nuvens de armazenamento de dados. Tendo isso em vista, este capítulo inicia contextualizando o modelo de nuvem computacional.

Na Seção 2.1 os principais conceitos sobre a computação em nuvem são introduzidos. Nela são apresentadas a definição, os principais benefícios, as características e os modelos da computação em nuvem.

Em seguida, na Seção 2.2, apresenta-se o modelo de armazenamento de dados em nuvem. Essa Seção inclui a definição de nuvem de armazenamento, os seus principais benefícios, a visão geral da arquitetura de nuvens de armazenamento e os seus principais casos de uso.

Outros elementos relevantes para este trabalho são os conceitos de deduplicação de dados e de federações de identidade. Isso porque o mecanismo proposto prevê a utilização dessas duas tecnologias na sua execução. Desta forma, este capitulo conta também com duas seções que apresentam o embasamento teórico de deduplicação de dados e de federações de identidade na Seção 2.3 e na Seção 2.4, respectivamente.

2.1 Computação em Nuvem

A computação em nuvem tornou-se amplamente adotada pelos diversos setores da sociedade seja para atender necessidades pessoais, comerciais ou acadêmicas

(FRAMINGHAM, 2011) (GORDON , HALE, *et al.*, 2012). Apesar da sua popularização, ela não apresenta uma única definição. Diferentes autores apresentam conceitos distintos para definir esse modelo computacional.

Armburst *et. al.* (2010) define a computação em nuvem como um modelo composto por duas partes: as aplicações oferecidas através da internet e a infraestrutura de *hardware* e *software* que prove o serviço. Por outro lado, Furht e Escalante (2010) definem a computação em nuvem como um modelo composto por um conjunto de características, formas de implantação, formas de oferta de serviço e de tecnologias especificas. Esteves (2011), por sua vez, apresenta um definição baseada na definição dos autores anteriores, e inclui uma classificação das características do modelo de nuvem. Através desta classificação, esse autor categoriza as características como variáveis ou comuns dos serviços em nuvem entre os serviços de nuvem.

Mediantes as diferentes definições, o Instituto Nacional de Padrões e Tecnologias dos Estados Unidos (*NIST*) apresentou no ano de 2011 um conceito que oferecesse uma perspectiva geral da computação em nuvem. Sob a autoria de Mell e Grance (2011), este conceito tornou-se o mais adotado como definição de computação em nuvem tornando-se amplamente utilizado tanto no contexto cientifico quanto no contexto empresarial. Mell e Grance (2011) conceituam a computação em nuvem, não pelos seus componentes, mas como um modelo composto de cinco características essenciais, três modelos de serviços e quatro modelos de implantação.

Mesmo havendo diferentes perspectivas conceituais, as nuvens computacionais destacam-se por apresentar um conjunto de benefícios inerentes ao modelo. Eles são, por exemplo, a diminuição de gastos com manutenção de infraestrutura, a oferta de serviço sob demanda, a alta disponibilidade dos serviços.

O Quadro 1 identifica e descreve alguns dos principais benefícios que a computação em nuvem oferece aos seus usuários.

Quadro 1. Benefícios da Computação em Nuvem.

BENEFÍCIO	DESCRIÇÃO

DIMINUIÇÃO DE CUSTOS (VISWANATHAN, 2012)	A computação em nuvem permite a oferta de serviço baseado no modelo "*pay as you use*". Nele o consumidor paga apenas pelos recursos e serviços que utilizar (sob demanda). Assim, cenários que envolvem altos investimentos, como por exemplo, a expansão de infraestrutura para o atendimento de demandas temporárias, podem ser substituídos pela adoção de recursos em nuvem. Além disso, ela proporciona a diminuição de custos relacionados a licenciamento, manutenção e *upgrade* de softwares.
FÁCIL ACESSO A INFORMAÇÃO (MELL E GRANCE, 2011)	Serviços em nuvem podem ser acessados independente de fuso horário ou da localização geográfica dos usuários. Para isso e há a necessidade apenas de conexão com a internet.
RÁPIDA IMPLANTAÇÃO DE SERVIÇOS (MELL E GRANCE, 2011)	Infraestruturas de nuvem oferecem vários recursos que facilitam a implantação de serviços. Esses recursos podem ser desde ambientes de armazenamento de dados configuráveis até máquinas virtuais preparadas para a hospedagem de aplicações.
CONFIABILIDADE, ESCALABILIDADE E DISPONIBILIDADE NA OFERTA DE SERVIÇOS (ARMBRUST *ET AL.*, 2010)	Infraestruturas de nuvem são ambientes que garantem que dados não serão perdidos e que haverá a mínima interrupção no acesso aos serviços. Isso é alcançado por meio da redundância de dados realizada nos servidores que compõe seus *data centers*.
BENEFÍCIO AMBIENTAL (MCKENDRICK, 2011)	Por meio da computação em nuvem, o consumo de energia dentro de organizações com refrigeração e equipamentos de fornecimento de energia são minímizados visto que toda infraestrutura pode ser terceirizada. As empresas não precisam hospedar equipamento interno e assim poupam nos custos de energia. Os provedores por sua vez

	apresentam data *centers* projetados especificamente para obter melhor eficiência energética.

Esses benefícios são comuns aos serviços em nuvem. Eles são obtidos devido a um conjunto de características que o modelo de nuvem computacional possui. Nas próximas seções o modelo de computação em nuvem e detalhado de acordo com suas características e formas de classificação.

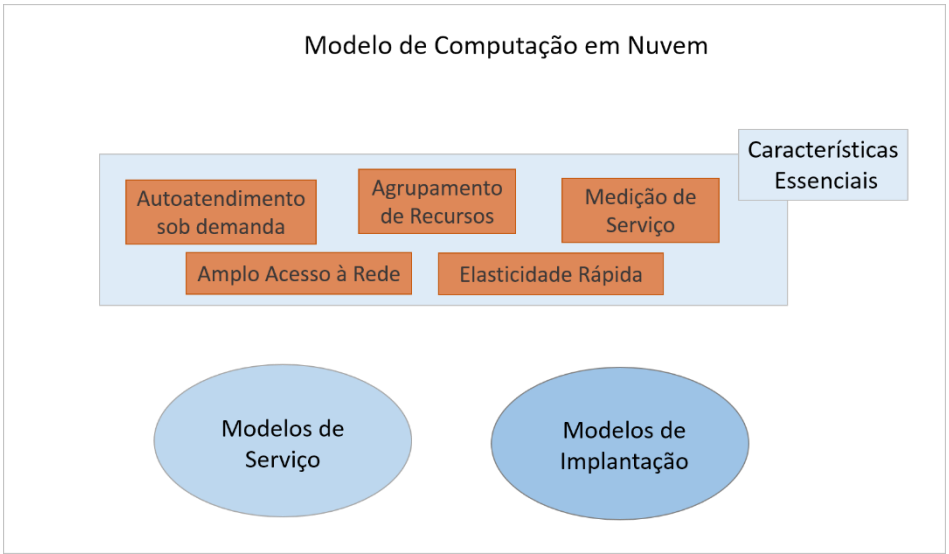
2.1.1 As Características da Computação em Nuvem

Conforme apresentado na seção anterior, o conceito de computação em nuvem de Mell e Grance (2011) é o mais adotado como definição de computação em nuvem. Esse modelo foi lançado como uma recomendação do Instituto Nacional de Padrões e Tecnologias dos Estados Unidos (*NIST*) e tem sido amplamente utilizado por diversos autores.

Conforme apresentado na seção anterior, a definição destes autores é composta por um conjunto de três componentes principais. O primeiro destes componentes são as características essenciais.

As características essenciais são os elementos que caracterizam a computação em nuvem e a torna diferente de outros modelos computacionais distribuídos. Mell e Grance (2011) definem cinco características essenciais: o autoatendimento sob demanda, o amplo acesso à rede, o agrupamento de recursos, a rápida elasticidade e a capacidade de medição do serviço. A Figura 2 apresenta uma visão geral das características essências da computação em nuvem.

Figura 1. Visão geral do modelo de Computação em Nuvem dos Autores Mell e Grance (2011) (Destaque das Características Essenciais).



Fonte: Elaborada pelo Autor.

O autoatendimento sob demanda está relacionado com a forma que os usuários obtêm recursos ou acessam serviços na nuvem. Nesse modelo o usuário obtém recursos computacionais de forma unilateral e automática sem a necessidade de interação humana com o provedor.

O amplo acesso à rede está relacionado com a forma que o serviço é oferecido. O serviço é ofertado por meio de redes computacionais e é acessado através de mecanismos padronizados, como *web services* que podem ser implementados por diversos tipos de plataformas, como dispositivos móveis e computadores *desktop*.

O agrupamento de recursos e a rápida elasticidade estão relacionados com a forma que o provedor oferece os recursos computacionais. Na primeira característica, os recursos são representados por servidores físicos e virtualizados que podem ser alocados dinamicamente de acordo com a necessidade do usuário. Eles são ofertados de forma transparente ao usuário, isto é, sem que ele tenha conhecimento exato da sua localização. A segunda refere-se à capacidade de adicionar ou remover recursos de forma automática a fim de proporcionar uma escalabilidade no serviço de acordo com a necessidade do consumidor. Essa característica é a base para o aparente provisionamento de recursos ilimitados e que podem ser oferecidos a qualquer momento.

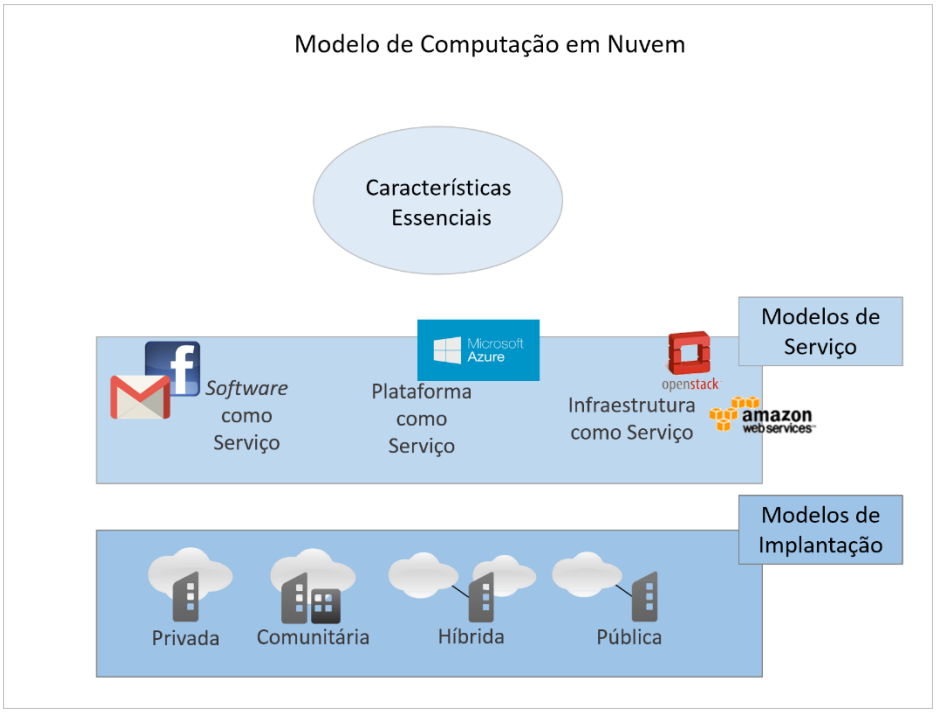
Por fim, a medição de serviço está relacionada com a capacidade de se obter métricas dos recursos e serviços disponibilizados pela nuvem. Elas permitem que tanto os usuários quanto o provedor possuam maior controle, transparência e capacidade de monitoramento dos recursos.

2.1.2 Os Modelos de Computação em Nuvem

Além das cinco características essenciais, Mell e Grance (2011) também apresentam duas formas de categorizar a computação em nuvem. Os modelos de computação em nuvem são categorizações que permitem classificar os diversos tipos de serviços ofertados em nuvem.

A primeira forma de categorização é de acordo com o modelo de serviços. Ela classifica a computação em nuvem de acordo com os tipos de serviços que oferece. A segunda forma é conforme modelos de implantação. Esta categoriza as nuvens de acordo com as formas que elas estão implantadas. A Figura 3 apresenta uma visão geral do modelo de serviço e de implantação apresentados por Mell e Grance (2011).

Figura 2. Visão geral do modelo de Computação em Nuvem dos Autores Mell e Grance (2011) (Destaque dos Modelos de Nuvem).



Fonte: Elaborada pelo Autor.

Os modelos de serviço definidos por Mell e Grance são três: o modelo de infraestrutura como serviço (*Infraestructure as a Service - IaaS*), de plataforma como serviço (*Plataform as a Service - PaaS*) e de software como serviço (*Software as a Service - SaaS*).

No modelo *IaaS* o provedor oferece recursos computacionais básicos ao consumidor. Nele o usuário pode utilizar a infraestrutura de processamento, armazenamento e redes do provedor. Essa infraestrutura pode ser ofertada por meio de componentes físicos (*hardware)* ou de forma virtualizada. Como exemplo de oferta de serviço segundo este modelo há os serviços oferecidos pelas empresas Amazon[6] e Rackpace[7]. Essas empresas disponibilizam desde espaços de armazenamento de dados até ambientes virtualizados aos seus usuários.

No modelo *PaaS* o provedor oferece um ambiente para a realização de todas as etapas relacionadas ao desenvolvimento de aplicações. Nesse modelo o usuário utiliza os recursos do provedor para desenvolver, hospedar, gerenciar e executar aplicações.

[6] https://aws.amazon.com
[7] https://www.rackspace.com

Como exemplos de oferta de serviços segundo este modelo tem se o Windows Azure Plataform[8] e o Google App Engine[9]. Ambos oferecem um ambiente para o desenvolvimento e a hospedagem de aplicações *web*.

No modelo *SaaS* o provedor oferece uma aplicação sendo executada em uma infraestrutura de nuvem. Os usuários utilizam os *softwares* disponibilizados acessando-os por meio de dispositivos clientes como navegadores *web*. O Facebook[10] e o Goggle Docs[11] são exemplos de serviços ofertados segundo o modelo *SaaS*.

Em relação aos modelos de implantação, Mell e Grance definem quatro formas de categorizar a computação em nuvem, elas são as nuvens públicas, as nuvens privadas, as nuvens comunitárias e as nuvens híbridas. As nuvens públicas são aquelas em que a infraestrutura de nuvem é estabelecida e controlada por uma organização terceirizada. Elas podem ser acessadas por um público geral e os dados armazenados nela não estão sob controle de seus usuários. As nuvens privadas, ao contrário do modelo anterior, são estabelecidas e controladas pela própria organização que a utiliza e, portanto, os dados armazenados nela estão sobre seu controle. As nuvens comunitárias são extensões das nuvens privadas diferenciando-se apenas por não pertencer a uma única organização, mas por um conjunto de organização que possuem um interesse em comum. O último modelo, isto é, a nuvem híbrida, é aquele que compreende mais de um dos modelos citados, isto é, compreende uma combinação dos modelos de nuvem pública, privada e comunitária.

2.2 As Nuvens de Armazenamento de Dados

Dentre os diversos tipos de recursos que podem ser oferecidos por meio da computação em nuvem, o armazenamento de dados tem recebido ampla adoção devido ao grande aumento da demanda por espaço de armazenamento (BORGMANN, 2012). O fator principal que impulsiona essa demanda está relacionado com o surgimento de novos

[8] http://azure.microsoft.com/
[9] https://appengine.google.com
[10] https://facebook.com
[11] https://docs.google.com/

dispositivos digitais capazes de acessar a internet. Esses dispositivos aumentam o tráfego de dados na rede e induzem a necessidade de armazená-los.

Nesse contexto, o armazenamento local de dados não é mais suficiente para atender as necessidades de alta capacidade armazenamento. É necessária a adoção de uma forma de armazenamento escalável e confiável que garanta aos proprietários o acesso eficiente a seus dados. O armazenamento de dados em nuvem surge então como um modelo que oferece uma solução para esse cenário.

O armazenamento de dados em nuvem consiste na oferta de espaço para armazenamento de dados em uma infraestrutura estabelecida de acordo com o modelo de computação em nuvem. Ele está inserido no modelo de serviço *IaaS* e é aderente a qualquer modelo de implantação de serviços em nuvem.

Todos os benefícios apresentados no Quadro 1 estão presentes no armazenamento em nuvem. Adicionalmente, alguns novos benefícios podem ser observados nesta forma de oferta de recurso. O Quadro 2 identifica e descreve esses benefícios.

Quadro 2. Benefícios do Armazenamento de Dados em Nuvem (BORGMANN, 2012)**.**

BENEFÍCIO	**DESCRIÇÃO**
OFERECE ESPAÇO DE ARMAZENAMENTO ILIMITADO	O armazenamento de dados em nuvem fornece a capacidade de armazenamento praticamente ilimitada. Assim, não há a necessidade do usuário se preocupar em ficar sem espaço de armazenamento ou de ter que aumentar a sua disponibilidade.
RECUPERAÇÃO E *BACKUP* DE DADOS	Dados armazenados em nuvem são mantidos em uma infraestrutura que garante sua disponibilidade e recuperação a qualquer momento pelos usuários.

OS DISPOSITOVOS DE ARMAZENAMENTO SÃO TRASPARENTES AO USUÁRIO	Todos os dispositivos utilizados para o armazenamento estão distantes fisicamente do usuário. Desta forma o usuário não deve se preocupar com tarefas relacionadas à instalação e configuração deles. Apesar disso, ele tem a sensação de estar armazenando dados em um dispositivo local.
PERMITE A SICRONIZAÇÃO DE DADOS	O armazenamento de dados em nuvem permite que usuários sejam capazes de sincronizar arquivos automaticamente em todos os seus dispositivos. Dessa forma, a versão mais recente de um arquivo salvo em seu computador pode está disponível em seu *smartphone*, por exemplo.
PERMITE O COMPARTILHAMENTO DE DADOS	O armazenamento de dados em nuvem permite que usuários sejam capazes de compartilhar arquivos facilmente com outros usuários, com um grupo específico ou até mesmo publicá-los para que todos possam vê-los na internet.

Os benefícios identificados no Quadro 1, juntamente com os definidos no Quadro 2, apresentam como o armazenamento de dados em nuvem é vantajoso quando comparado com outras de formas de armazenamento de dados.

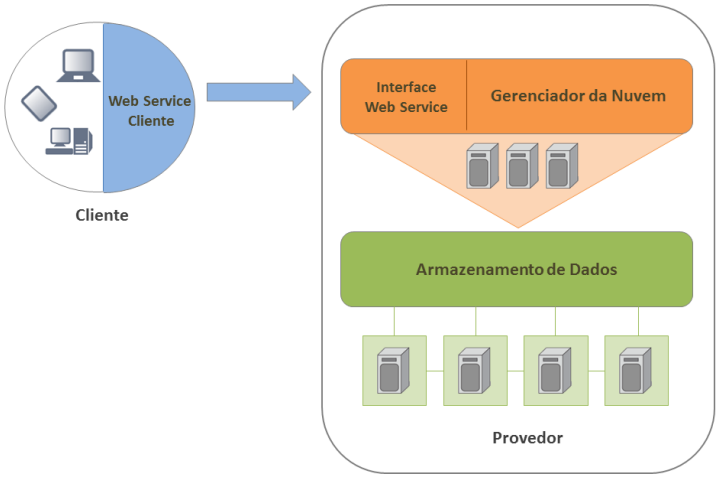
2.2.1 Visão Geral de Serviços de Armazenamento em Nuvem

As nuvens de armazenamento são serviços que possuem a infraestrutura de armazenamento e permitem que clientes possam armazenar dados e recuperá-los do

ambiente oferecido. Borgmann *et al.* (2012) apresenta esses serviços como “uma rede de centros de dados distribuídos que normalmente usa tecnologias de computação em nuvem, como virtualização, e oferece algum tipo de interface para armazenar dados”.

A Figura 4 apresenta uma visão geral da arquitetura de serviços de armazenamento em nuvem.

Figura 3. Visão Geral da Arquitetura de Serviços de Armazenamento em Nuvem.



Fonte: Elaborada pelo Autor.

Um serviço de armazenamento em nuvem, normalmente, pode ser dividido em duas perspectivas diferentes: a do cliente e a do provedor.

O cliente refere-se ao consumidor do serviço e é representado por aplicações que acessam a nuvem. Elas podem ser executadas em diversos tipos de dispositivos como computadores *desktop* e dispositivos móveis e variam desde sistemas *web* até aplicativos de dispositivos móveis. Essas aplicações implementam *web service* cliente, isto é, que realizam requisições, e realizam o acesso à infraestrutura do provedor por meio deles.

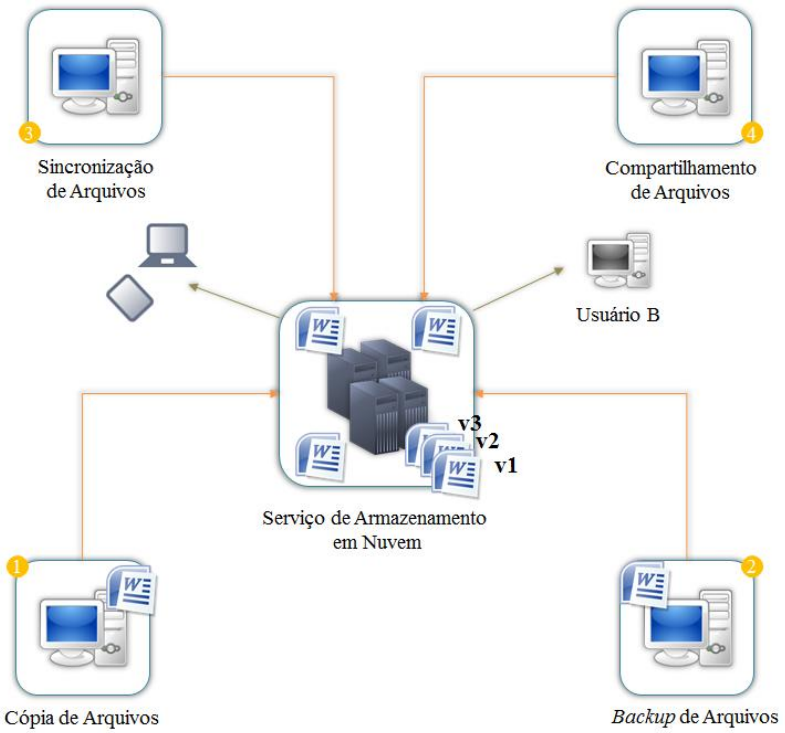
O provedor é o responsável por fornecer e administrar o serviço de armazenamento em nuvem. Ele apresenta dois componentes principais em sua infraestrutura: os servidores gerenciadores da nuvem e os servidores de armazenamento de dados. Os servidores gerenciadores são aqueles responsáveis por receber as requisições dos clientes. Eles apresentam interfaces *web services* que recebem as requisições e as

transformam em uma operação sobre a infraestrutura, como por exemplo, o envio ou a obtenção de arquivos. Os servidores de armazenamento são os responsáveis por hospedar os arquivos enviados e disponibilizá-los quando solicitado pelo gerenciador.

2.2.2 Casos de Uso dos Serviços de Armazenamento em Nuvem

Borgmann *et al.* (2012) apresenta em seu trabalho os principais cenários de utilização relacionado com os serviços de armazenamento em nuvem. Os quatro cenários de utilização definidos pelo autor são: a cópia, o *backup*, a sincronização e o compartilhamento de arquivos. A Figura 5 apresenta cada um deles.

Figura 4. Casos de Uso dos Serviços de Armazenamento em Nuvem.



Fonte: Elaborada pelo Autor.

A cópia de arquivos (item 1 da Figura 5) consiste na capacidade de armazenar arquivos remotamente e poder resgatá-los a partir de diversos tipos de dispositivos,

como celulares ou computadores *desktop*. Este cenário é conveniente, por exemplo, para usuários que querem obter um arquivo alterado em outras localidades diferente. Essa funcionalidade e facilmente verificada em serviços de armazenamento em nuvem como Dropbox (2015), Onedrive (2015) e Google Drive (2015) que permitem o envio de arquivos à nuvem por meio de seus aplicativos clientes.

O *backup* (item 2 da Figura 5) de arquivos está relacionado com a capacidade de manter em nuvem versões de um mesmo arquivo de forma que possam ser acessadas posteriormente. Esta funcionalidade é importante para usuários que após armazenarem diversas vezes um mesmo arquivo na nuvem, desejam obter versões anteriores do mesmo. A aplicação *web* do Dropbox fornece um exemplo de tipo de caso de uso. Nela é possível resgatar versões de arquivos armazenados que já foram removidos da nuvem pelo usuário.

A sincronização (item 3 da Figura 35) consiste na capacidade de configurar diferentes dispositivos clientes para obter automaticamente ou não um mesmo arquivo de forma consistente e atualizada. Neste caso de uso um usuário que costuma realizar constantes modificações em dispositivos diferentes, obtém em cada um deles, uma versão atualizada de seu arquivo. Para isto basta apenas configurar os dispositivos para sincronizarem os arquivos. Nos clientes *desktop* dos serviços Dropbox, Skydrive e Google Drive é possível verificar essa funcionalidade. Neles um diretório é criado no dispositivo do usuário e este é monitorado a fim de manter os arquivos sincronizados na nuvem.

Por fim o compartilhamento de arquivos (item 4 da Figura 5) refere-se capacidade de atribuir a usuários diferentes a capacidade de acessar um mesmo arquivo armazenado na nuvem. Esse cenário é conveniente quando um usuário deseja compartilhar um arquivo com outro ou deseja publicá-lo ao público em geral. O Google Drive é um exemplo de serviço que se destaca na oferta dessa funcionalidade. Ele permite que arquivos em nuvem sejam compartilhados com outros usuários do serviço ou por meio de uma URL pública para o público em geral. Além disso, Ele possui uma integração com o serviço de e-mail do Google que permite que arquivos de qualquer tamanho hospedados na nuvem sejam compartilhados como anexo de e-mails.

2.3 A Deduplicação de Dados

O surgimento de serviços de armazenamento de grandes quantidades de dados gerou a necessidade da criação de técnicas que viabilizassem uma melhor gerência do volume de informação hospedada por eles. Desafios relacionados com a forma mais eficiente de armazenar e de aproveitar a banda de transmissão desses dados tornam-se crítico na medida em que os usuários armazenam cada vez mais arquivos (LOWE, 2012).

A deduplicação de dados é uma técnica desenvolvida para o tratamento mais eficiente de dados armazenados. Através dela a redundância de informação é identificada e ao invés de serem armazenados dados iguais nos servidores, apenas uma única instância deles é armazenada. Todos os usuários que possuem uma cópia dessa instância mantêm uma referência para ela (DEEPAK e SHARMA, 2013).

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 (D., KAISER, *et al.*, 2012) (JIN e MILLER, 2009).

Em nuvens de armazenamento de dados é comum o acumulo de vários arquivos iguais na infraestrutura do serviço. De forma a resolver tal problema, as nuvens têm aderido à utilização da técnica de deduplicação para aprimorar a forma que os dados de seus usuários são armazenados e em alguns casos, para otimizar o uso da largura de banda na comunicação entre o usuário e a nuvem. As nuvens do Dropbox (2015), Mozy (2015) e Memopal (2015) são exemplos de serviços de armazenamento que usufruem de uma técnica de deduplicação de dados.

A seguir, serão apresentadas as principais características e diferenças dos métodos de implementação de deduplicação de dados.

2.3.1 Os Métodos de Implementação de Deduplicação de Dados

A deduplicação de dados, conforme introduzido na Seção 2.3, é uma técnica em que dados iguais são identificados e tratados a fim de que apenas uma instância da informação seja armazenada nos servidores dos serviços de armazenamento de dados. Essa técnica apresenta formas de realização distintas. Essas formas, denominadas de métodos de implementação, introduzem diferentes vantagens e desvantagens ao serviço que as utiliza.

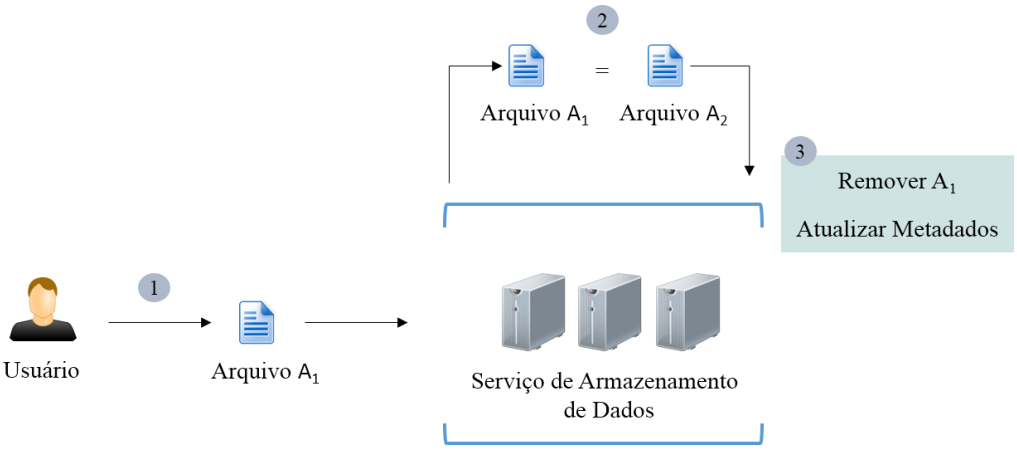
Os métodos de implementação de deduplicação de dados podem ser categorizadas de acordo com duas classificações. São elas: quanto à granularidade de aplicação da técnica e quanto ao local onde a técnica será executada.

Na primeira classificação considera-se a forma operacional da técnica de deduplicação sobre o arquivo. De forma geral, duas estratégias principais podem ser utilizadas, são elas: a deduplicação em nível de arquivo e a deduplicação em nível de blocos (DEEPAK e SHARMA, 2013) (MEISTER e BRINKMANN, 2009). As Figuras 6 e 7 apresentam uma visão geral desses métodos de implementação.

Na deduplicação em nível de arquivo o mecanismo considera todo o arquivo para identificar dados redundantes (etapa 1). Um arquivo enviado é comparado com outros armazenados nos servidores (etapa 2). Quando uma cópia é encontrada ela é marcada como duplicada e não é armazenada. Então, a base de metadados é atualizada com as referências de localização necessárias, como por exemplo, o caminho que o arquivo enviado deveria ser armazenado (etapa 3). A verificação de igualdade entre arquivos normalmente é realizada calculando-se os *hash* dos dados e comparando-os.

Essa estratégia, apesar viabilizar a verificação de redundância apresenta a limitação de que pequenas modificações no mesmo arquivo não são identificadas pela deduplicação. Mesmo que apenas um único byte seja modificado, nada é aproveitado entre eles, pois os valores *hash* utilizados para compara-los serão diferentes.

Figura 5. Método de Implementação de Deduplicação em Nível de Arquivo.

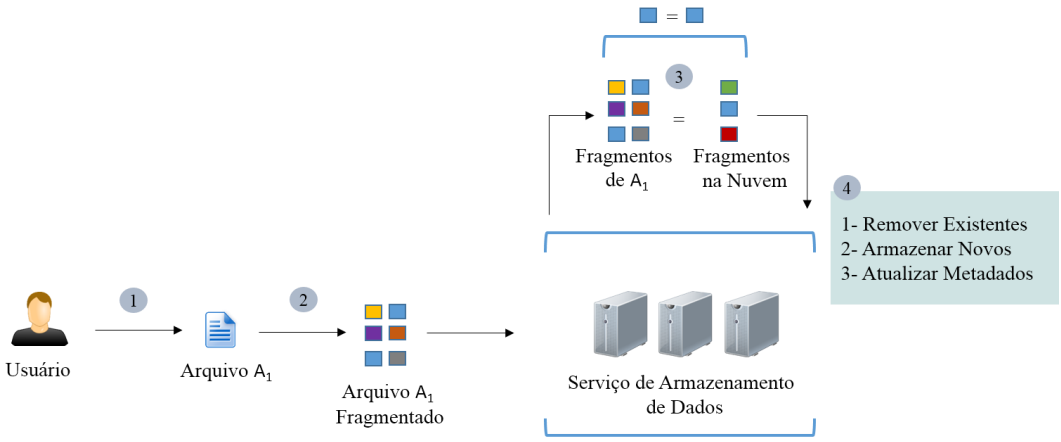


Fonte: Elaborada Pelo Autor

A deduplicação em nível de blocos é similar à que ocorre em nível de arquivos, pois a comparação ainda é realizada através do *hash* dos dados. Porém, nessa estratégia, o arquivo enviado é fragmentado antes de ser realizada a verificação de redundância (etapas 1 e 2). Deste modo, durante a deduplicação, os blocos de arquivos recebidos serão comparados com outros blocos armazenados nos servidores do serviço (etapa 3). Então, apenas blocos novos serão armazenados e todos os blocos que correspondem ao enviado serão referenciados para o usuário dono do arquivo enviado.

Neste método a eficiência na detecção de dados iguais é aprimorada, visto que blocos iguais entre 2 arquivos que apresentam pequenas diferenças são detectados e deduplicados. Contudo, o processamento realizado durante o processo de deduplicação e durante o momento de disponibilizar o arquivo ao usuário (ex. solicitação de *download*) é maior exigindo maior tempo de processamento para ser realizado.

Figura 6. Método de Implementação de Deduplicação em Nível de Blocos.



Fonte: Elaborada Pelo Autor

Ainda no método de implementação de deduplicação em nível de blocos é possível categoriza-lo de acordo com a forma que os blocos são deduplicados. As categorias são duas: deduplicação de blocos estáticos ou deduplicação de blocos baseada em conteúdo do bloco (DEEPAK e SHARMA, 2013) (MEISTER e BRINKMANN, 2009) (BO e LI, 2013).

A deduplicação de blocos estáticos considera que o arquivo é divido em blocos de tamanhos fixos e, portanto, os limites de cada fragmento são definidos em posições específicas do arquivo, como por exemplo, blocos de 64bytes.

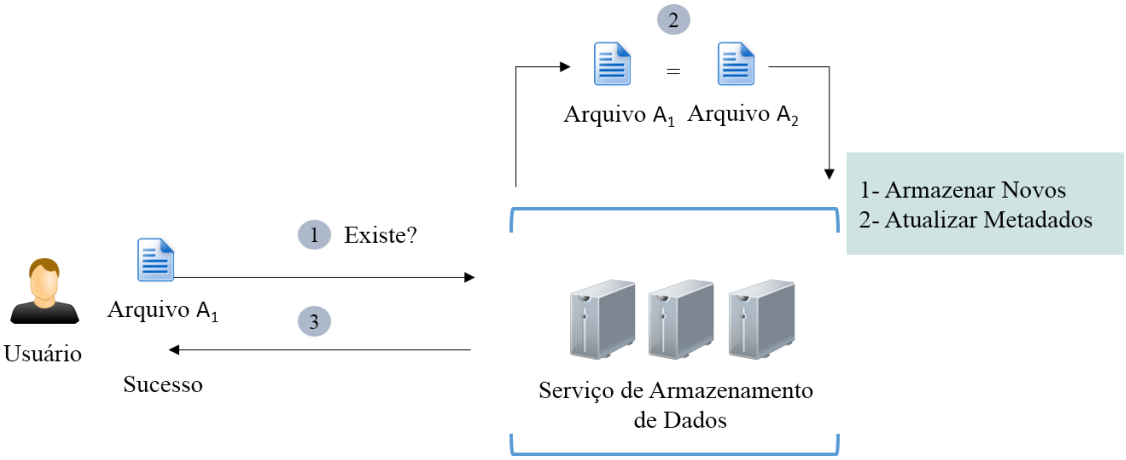
Apesar de apresentar melhor eficiência ao verificar dados repetidos, pois não realiza nenhum pré-processamento antes ou durante a deduplicação, ela apresenta baixa eficácia. Isso porque, está técnica não é capaz de detectar dados iguais mediante pequenos deslocamentos de informação dentro de seu conteúdo.

Por outro lado, a técnica de deduplicação baseada em conteúdo de blocos define o tamanho de bloco através do seu conteúdo e não na posição dos dados do arquivo. Essa técnica apresenta maior eficiência, pois é capaz de detectar blocos de dados iguais mesmo mediante modificações dentro do arquivo. Apesar disso, ela é uma técnica menos eficiente devido a necessidade de realizar um pré-processamento durante o processo de deduplicação.

A segunda classificação considera o local onde a deduplicação de dados ocorre. Dois modelos existem nessa classificação: o de deduplicação na fonte (*source based*

deduplication) e o de deduplicação no destino (*target based deduplication*). A Figura 8 e a Figura 9 apresentam uma visão geral dessas estratégias.

Figura 7. Método de Implementação de Deduplicação na Fonte.

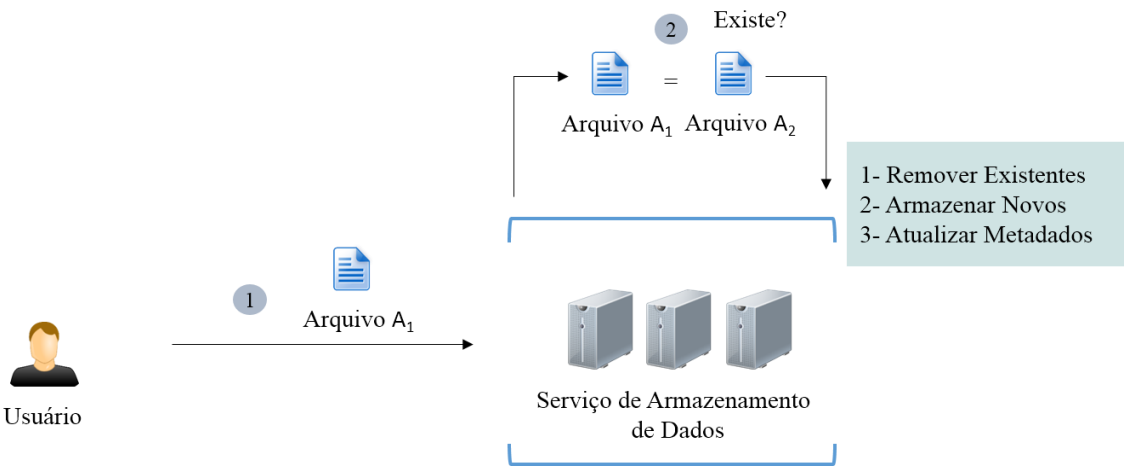


Fonte: Elaborada Pelo Autor

No método de deduplicação na fonte a técnica é aplicada antes dos dados serem enviados para o serviço de armazenamento de dados. O cliente calcula o valor *hash* dos dados e os envia para o servidor (etapa 1). O servidor por sua vez utiliza esses *hash* para identificar redundância de informação (etapa 2). Ao final, ele solicita apenas os dados cujos *hash* não foram encontrados na sua infraestrutura. Caso o arquivo já exista, o serviço apenas atualiza sua base dados com as referências necessárias.

Como os dados já armazenados pelo serviço não são enviados, então tanto armazenamento quando o uso da largura de banda é otimizado nesse modelo. Apesar disso, esse modelo introduz um processamento adicional em cada cliente exigindo maior poder computacional deles.

Figura 8. Método de Implementação de Deduplicação no Destino.



Fonte: Elaborada Pelo Autor

No método de deduplicação no destino os dados do usuário são sempre enviados ao servidor, sendo eles redundantes ou não. O cliente não percebe que seus dados estão sendo deduplicados, e, portanto o mecanismo não impacta no uso de seus recursos computacionais. O servidor armazena os dados do cliente caso não seja encontrado uma copia dele. Caso uma cópia exista, é mantida apenas uma cópia temporária dos dados até a deduplicação ser executada.

A deduplicação no destino apresenta o benefício de permitir a realização de deduplicação como uma tarefa de segundo plano e pode ser executada em momento oportunos para o serviço (ex. momento de baixo acesso pelos clientes). Por outro lado, esse método exige que o serviço execute todo o processamento da deduplicação e, portanto, apresenta o uso de CPU muito mais intensivo durante o processo de deduplicação.

2.4 As Federações de Gestão de Identidade

A formação de redes organizacionais colaborativas tem se tornado cada vez mais comum no cenário global. Essas redes consistem em organizações distribuídas geograficamente que possuem um ou mais serviços de interesse em comum e operam de

forma colaborativa pelas redes de computadores. As CNOs (*Collaborative Networks of Organizations*) e NRENS (*National Research and Educational Networks*) são exemplos de tais redes.

Nas redes colaborativas, o aumento da oferta de serviços e recursos para seus usuários motivam a formação de associações denominadas federações. De acordo com Wangham *et al.* (2010), as federações são:

> uma forma de associação de parceiros de uma rede colaborativa que usa um conjunto comum de atributos, práticas e políticas para trocar informações e compartilhar serviços, possibilitando a cooperação e transações entre os membros da federação

Nesse contexto, uma federação de gestão de identidade (*Federated Indentity Management - FIM*) é representada por um conjunto de organizações parceiras que apresentam o serviço de gestão de identidade compartilhado entre si. A gestão de identidade consiste em um sistema que agrega políticas, processos e tecnologias que permitem os parceiros da federação manter e manipular os dados que representam a identidade dos seus usuários. Como exemplos de tais federações tem-se a *Canadian Access Federation*, a Edugate e a CAFe brasileira (RNP, 2015).

Existem diversas soluções para implantações de FIMs. O Shibolleth[12], o OpenID[13] e o OAuth[14] são exemplos que podem ser utilizados para a criação de serviços de federações de identidade.

Na perspectiva do usuário, um dos principais benefícios oferecidos pelas FIMs é a função de autenticação única (*Single Sign-On* – SSO). O SSO permite ao usuário autenticar-se apenas uma vez no domínio da federação e obter credencias válidas para todos os seus serviços. Desta forma o usuário possui a facilidade de realizar o processo de autenticação uma única vez enquanto usufrui dos recursos e serviços da federação. Além disso, ele não precisa gerenciar diferentes senhas para cada serviço, necessitando memorizar apenas uma senha de acesso. Exemplos de soluções que permitem a implantação de um serviço SSO são o Shibolleth e o OpenID.

Além da autenticação, as federações de identidade também são capazes de oferecer o serviço de autorização sobre dados dos usuários. Um dos principais

[12] https://shibboleth.net/
[13] http://openid.net/
[14] http://oauth.net/

representantes nesse contexto é o *framework* OAuth, amplamente utilizado pelas redes sociais para fornecer um mecanismo de autorização de acesso aos dados dos seus usuários como por exemplo, listas de amigos e dados de perfil.

Tendo em vista os conceitos introduzidos, nas próximas seções serão apresentados os componentes e o funcionamento de uma federação de gestão de identidade.

2.4.1 Os Componentes de Federações de Identidade

Conforme introduzido na seção anterior, existem diferentes soluções para implantações de federações de identidade. Essas soluções, normalmente, são formadas por dois componentes principais.

O primeiro é o componente de *software*. Ele é representado por dois elementos principais: os provedores de identidade (*Identity Providers -IdPs*) e os provedores de serviços (*Service Providers - SPs*). Os provedores de identidade são responsáveis por manter e emitir as credenciais dos usuários. Essas credenciais representam a identidade do usuário. Ela consiste em um identificador e um conjunto de atributos que o representam. Os provedores de serviço, por sua vez, oferecem recursos aos usuários que possuem a sua credencial emitida por um IdP. O usuário apenas acessará recursos após ter a autenticidade de sua identidade verificada e ter comprovado que possui atributos para acessá-los.

O segundo componente é representado pelo protocolo de funcionamento. Eles são utilizados para definir o padrão de interações entre os componentes de *software* entre si e entre eles e o usuário. Um dos principais exemplos desses protocolos é o padrão SAML (*Security Assertion Markup Language*) especificado pelo Comitê Técnico de Serviços de Segurança da Organização para o Avanço dos Padrões de Informações Estruturadas (*Organization for the Advancement of Structured Information Standards - OASIS*).

Tais componentes formam um sistema de gestão de identidade que permite prover o serviço de autenticação e de gerenciamento de atributos dos usuários

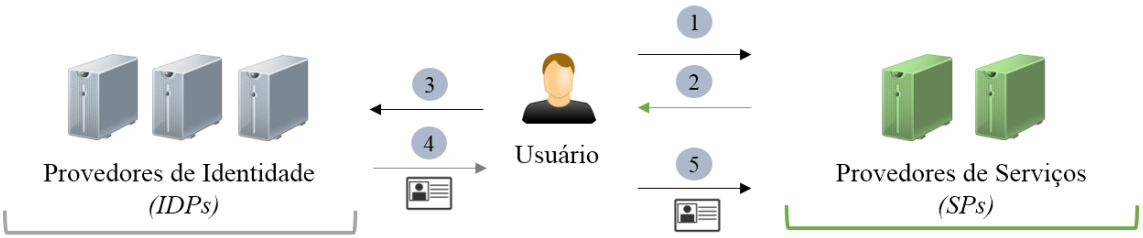
(WANGHAM *et al.*, 2010).

2.4.2 O Funcionamento de Federações de Identidade

Nesta seção é apresentado como ocorre a interação entre os componentes apresentados na Seção 2.4.1 e o usuário a fim de prover o serviço de gestão de identidade. Nesta descrição o funcionamento é apresentado desconsiderando elementos presentes na utilização de protocolos específicos, como por exemplo, formato de mensagem.

A Figura 10 apresenta uma visão geral de funcionamento de uma FIM.

Figura 9. Modelo de Gestão de Identidade Federado.



Fonte: Elaboração Própria.

Conforme apresentado na imagem, o início do processo de emissão de credenciais ocorre entre o usuário e o SP da federação. O usuário normalmente entra na página do serviço que deseja utilizar e se identificar para acessar seus recursos (etapa 1). O provedor de serviço, então, informa ao usuário os IdPs disponíveis para autentica-lo (etapa 2).

Tendo conhecimento do (s) IdP(s), o usuário seleciona aquele em que deseja se autenticar (etapa 3). Nele o usuário insere os dados de autenticação que, normalmente, são seu *login* e senha, e, então, solicita a emissão de suas credenciais. Caso o usuário esteja cadastrado na federação, ele receberá suas credencias de acesso. Essa credencial contêm o identificador e os atributos do usuário (etapa 4.)

Por fim, o usuário tendo posse de sua credencial, acessa o provedor de serviço fornecendo-a em suas requisições por recursos (etapa 5).

3. UM NOVO MECANISMO DE COMPARTILHAMENTO DE DADOS

Conforme apresentado no Capítulo 1, esta pesquisa propõe um novo mecanismo de compartilhamento de dados para nuvens de armazenamento. O mecanismo é denominado de *compartilhamento por caixas de interesses*. Ele tem o objetivo de viabilizar a troca de arquivos entre quaisquer usuários que façam parte de um serviço de armazenamento de dados em nuvem desde que apresentem potenciais interesses em comum.

O compartilhamento por caixas de interesses baseia-se na utilização de duas tecnologias auxiliares. A primeira é a deduplicação de arquivos no destino (ver Seção 2.3). Essa tecnologia apoia o mecanismo no processo de identificação de usuários com interesses em comum. A segunda é um sistema de gestão de identidade federada (ver Seção 2.4). No mecanismo proposto, ele atua como um centro provedor de atributos. Tais atributos são utilizados pelos usuários para controlar o acesso sobre seus arquivos compartilhados.

Considerando o referencial teórico do Capitulo 2, este capítulo apresenta e descreve o compartilhamento por caixas de interesses. Inicialmente, é introduzida uma visão geral na Seção 3.1. Nesta seção, são apresentadas as principais definições, conceitos e componentes do mecanismo. Em seguida, na Seção 3.2, o funcionamento do mecanismo é detalhado. Por fim, na Seção 3.3, apresenta-se o protótipo desenvolvido durante esta pesquisa. Seus módulos e a forma que ele implementa o mecanismo são descritos nesta última seção.

3.1 Visão Geral

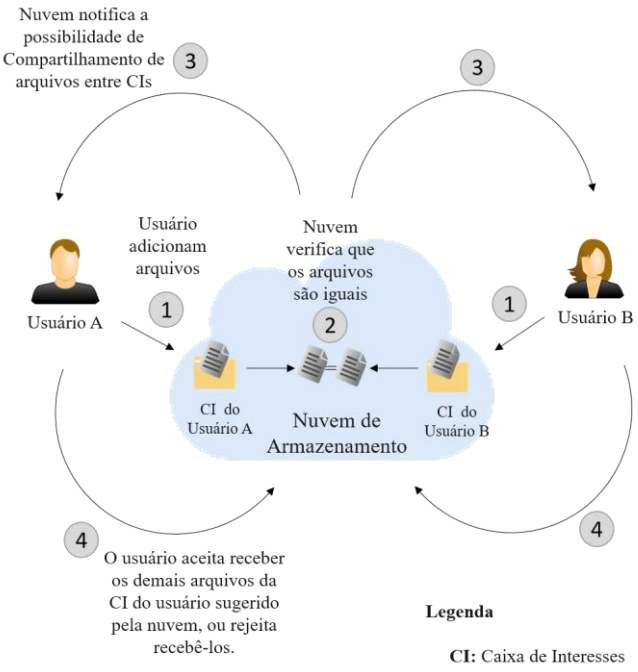
O compartilhamento por caixas de interesses consiste em um mecanismo de compartilhamento de dados em nuvem. Seu objetivo é promover a colaboração entre quaisquer usuários do serviço de armazenamento de dados em nuvem, isto é, tanto os que se conhecem quanto os que não se conhecem, através da troca de arquivos.

Na perspectiva do usuário, o mecanismo é realizado a partir de estruturas denominadas de caixas de interesses. Através delas os usuários agrupam arquivos que desejam compartilhar com outros usuários.

O serviço de armazenamento de dados em nuvem, por sua vez, monitora essas estruturas e verifica quais usuários possuem potenciais interesses em comum. Essa verificação é baseada na igualdade de arquivos armazenados pelos usuários. Aqueles que apresentam arquivos em comum são identificados como indivíduos com potenciais interesses em comum. Usuários com potenciais interesses em comum são, então, suscetíveis a realizar o compartilhamento de arquivos entre si.

A Figura 11 apresenta uma visão geral do mecanismo de compartilhamento por caixas de interesses.

Figura 10. Visão Geral do Mecanismo de Compartilhamento por Caixas de Interesses.



Fonte: Elaboração Própria.

Conforme pode ser observado na Figura 11 apresenta os usuários adicionam seus arquivos nas caixas de interesses (item 1). Então, o serviço de armazenamento de dados em nuvem identifica quando os usuários possuem uma ou mais caixas de interesses com arquivos em comum em relação a outras caixas de interesses de outros usuários da nuvem (item 2). Quando a nuvem identifica as caixas de interesses com arquivos em comum, ela notifica os usuários dessas caixas sobre a possibilidade de compartilhamento de dados uns com os outros (item 3). O usuário, então, pode aceitar ou não compartilhar os demais arquivos de sua respectiva caixa de interesses com o usuário sugerido pela nuvem.

O mecanismo conta também com um componente responsável por fornecer meios para o usuário controlar com quem deseja compartilhar seus dados mesmo que o receptor do compartilhamento seja uma pessoa desconhecida. Este componente é denominado de centro provedor de atributos e é representado por um sistema de gestão de identidade federada.

Nas Seções seguintes, são apresentados o conceito de caixas de interesses, o processo de identificação de arquivos iguais entre usuários diferentes, a atuação do provedor de atributos e uma visão geral de funcionamento do mecanismo.

3.1.1 As Caixas de Interesses

Conforme apresentado na Seção 3.1 as caixas de interesses são elementos fundamentais para a realização do mecanismo. Elas armazenam arquivos que poderão ser compartilhados na nuvem.

As caixas de interesses são representadas por pastas nos serviços de armazenamento de dados em nuvem. Elas se distinguem das pastas comuns 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 sua data de criação. Eles são definidos assim que o usuário cria sua caixa. Os metadados específicos são os

atributos que o usuário obtém a partir de um centro provedor de atributos. Eles, diferentemente dos metadados gerais, podem ser definidos pelo usuário a qualquer momento enquanto sua caixa de interesse existir.

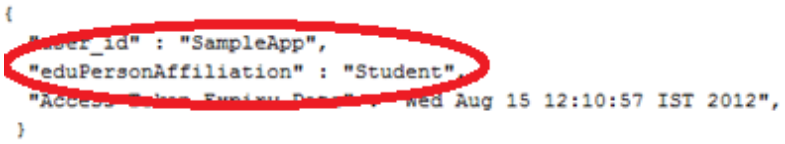
A Figura 12 e a Figura 13 apresentam exemplos de atributos disponibilizados por sistemas de gestão de identidade federada. No exemplo, o atributo o atributo destacado possui no *eduPersonAffiliation* e valor *Student*.

Figura 11. Exemplo de Atributo de Federação de Identidade (Exemplo Simplificado do Protocolo SAML).

```
<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>
```

Fonte: Elaborada pelo Autor

Figura 12. Exemplo de Atributo de Federação de Identidade (Exemplo Simplificado do Protocolo OAuth).



Fonte: Elaborada pelo Autor

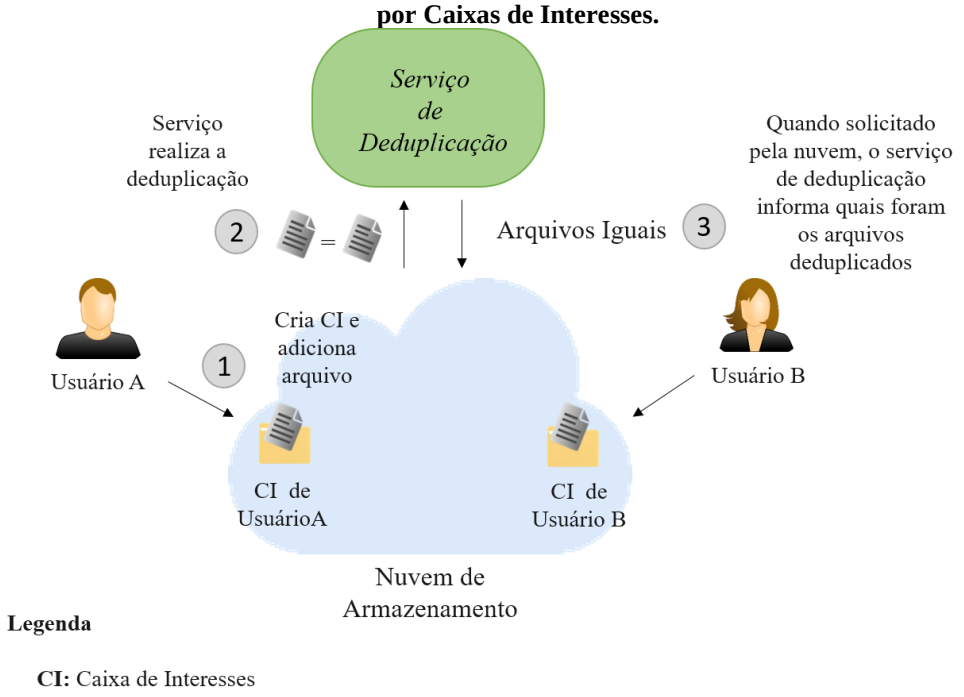
Nas caixas de interesses, cada metadado é representado por um par no formato <*propriedade: valor*>. Nos casos em que um usuário possui mais de um valor para um mesmo atributo (atributo multivalorado), considera-se um par para cada valor, isto é, os pares $< propriedade = valor_i > e < propriedade = valor_{i+1} >$ indicam que $propriedade$ possui os valores $Pval_i$ e $Pval_{i+1}$.

Portanto, para fins de exemplificação, uma caixa de interesse que possui um metado específico obtido a partir do atributo ilustrado nas Figura 12 e Figura 13, possui o par $< eduPersonAffiliation = student >$ como metadado.

3.1.2 A Identificação de Arquivos Iguais entre Usuários Diferentes

A identificação de arquivos iguais é uma das principais operações realizadas pelo mecanismo de compartilhamento por caixas de interesses. Por meio dela, a nuvem é capaz de identificar usuários que podem compartilhar dados entre si. O mecanismo utiliza as caixas de interesses e a tecnologia de deduplicação de dados da nuvem para identificar arquivos iguais entre usuários diferentes. A Figura 14 apresenta uma visão geral desse processo.

Figura 13. Visão Geral de Identificação de Arquivos pelo Mecanismo de Compartilhamento por Caixas de Interesses.



Fonte: Elaboração Própria.

Conforme ilustra a Figura 14, apenas arquivos armazenados em caixas de interesses são passiveis de compartilhamento. A deduplicação de dados, por sua vez, é utilizada no mecanismo para a identificação de arquivos iguais armazenados nas caixas de interesses. O serviço de deduplicação atua normalmente verificando os dados redundantes (item 2), contudo, sempre que solicitado, ele notifica o serviço de nuvem sobre a existência desses dados (item 3). Através dessas notificações, o serviço de

armazenamento identifica arquivos em comum entre usuários que desejam compartilhar arquivos por meio do mecanismo, isto é, arquivos em comum armazenados em caixa(s) de interesses dos usuários.

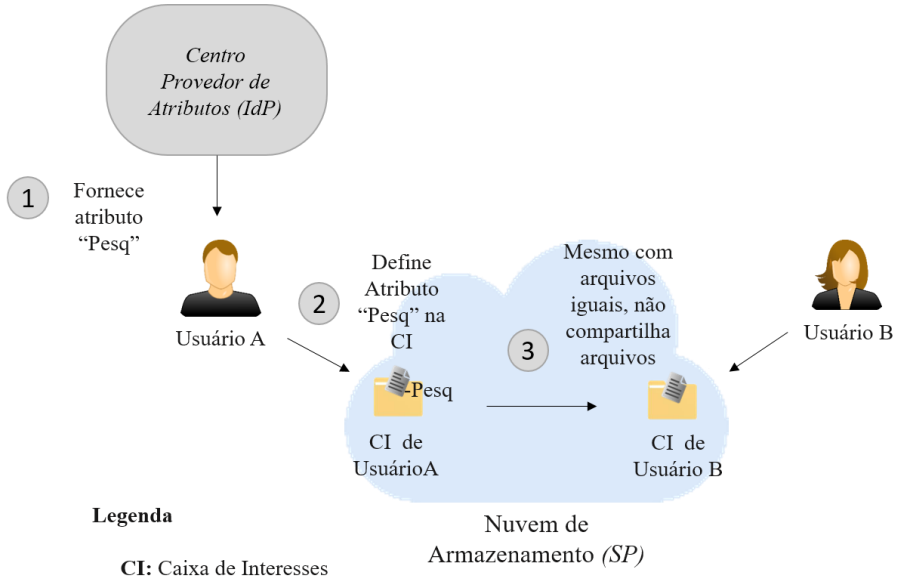
3.1.3 O Provedor de Atributos

O mecanismo de compartilhamento por caixas de interesses conta também com um recurso que fornece meios para o usuário controlar com quem deseja compartilhar seus dados mesmo que o receptor do compartilhamento seja uma pessoa desconhecida. Este componente é denominado de centro provedor de atributos e é representado por um sistema de gestão de identidade federada.

O centro provedor de atributos consiste, especificamente, no provedor de identidade (*Identity Provider - IdP*) do sistema de gestão de identidade federada. Conforme apresentado na Seção 2.4, os IdPs disponibilizam as informações de identidade dos usuários aos provedores de serviços da federação (*Service Provider - SP*). No contexto deste mecanismo a nuvem de armazenamento é o serviço que consome as credenciais e que oferece recursos aos usuários ex. (*upload* e *download* de arquivos e criação de diretórios). Portanto, ela representa o *SP* da federação de identidade

A Figura 15 apresenta uma visão geral da participação do centro provedor de atributos e da utilização dos atributos dos usuários no mecanismo.

Figura 14. Visão Geral do Provedor de Atributos e a Utilização dos Atributos dos Usuários.



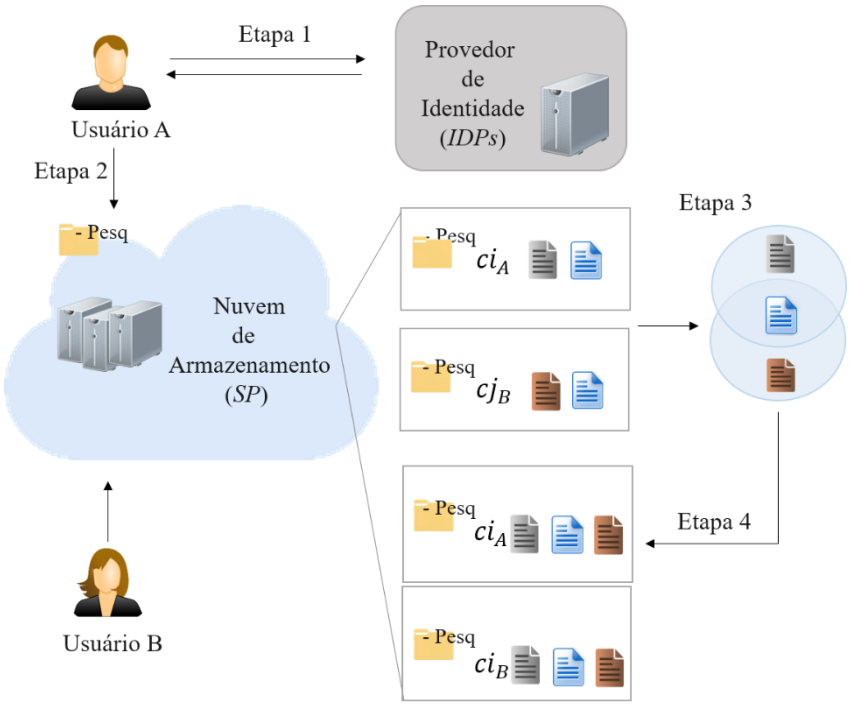
Fonte: Elaboração Própria.

O centro provedor de atributos atua fornecendo atributos para os usuários do serviço de armazenamento de dados em nuvem (item 1). Os usuários podem definir esses atributos nas caixas de interesses e restringir o compartilhamento dos arquivos na sua caixa (item 2). O serviço de armazenamento de dados em nuvem, por sua vez, apenas identificará outros usuários que possuam caixas de interesses com os mesmos atributos. Desta forma, usuários com caixas de interesses que possuem arquivos iguais, mas possuem atributos diferentes não são identificados como usuários com potencial interesses em comum e, portanto, não poderão compartilhar dados entre si.

3.1.4 Visão Geral do Mecanismo

Conforme apresentado a cima, o mecanismo de compartilhamento por caixas de interesses é realizado por meio da interação de três elementos principais: a nuvem de armazenamento, seus usuários e o centro provedor de atributos. Esses componentes são apresentados na Figura 16 abaixo.

Figura 15. Visão Geral do Funcionamento de Compartilhamento de Dados por Caixas de Interesses.



Fonte: Elaborada pelo Autor

Conforme se observa, o mecanismo ocorre por meio de quatro etapas principais. Elas 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. Esses atributos são fornecidos através da credencial emitida pelo provedor de atributos (IdP). O processo de obtenção da credencial ocorre conforme o modelo de gestão de identidade apresentado na Seção 2.4. Após esta etapa o usuário possuirá o acesso à nuvem de armazenamento de dados e aos seus atributos.

Na segunda etapa o usuário cria sua caixa de interesses. Ele então deve adicionar arquivos nela a fim de promover o compartilhamento. Ao adicioná-los, o serviço de deduplicação realiza a procurar por arquivos redundantes. Ao encontrar arquivos duplicados ele registra essa informação para fornece-la posteriormente a nuvem quando solicitado.

Para controlar a disponibilização dos dados compartilhados, o usuário deve incluir os metadados específicos da caixa de interesse. Desta forma, apenas as caixas que possuírem 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.

A identificação de usuários com potenciais interesses em comum ocorre na terceira etapa. Para isso, a nuvem conta com um componente denominado de Gerenciador de Caixas de Interesses. Esse componente é responsável por solicitar os registros de arquivos duplicados ao serviço de deduplicação. Após receber essas informações, o componente gerenciador de caixas de interesses, procura pelos usuários que possuem caixas de interesse com arquivos iguais.

Por fim, a nuvem promove o compartilhamento dos arquivos na etapa 4. Ela sugere o compartilhamento entre os usuários após identificar quais possuem arquivos iguais em caixas de interesses com mesmos atributos. 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.

3.2 Descrição do Funcionamento

Nesta seção são descritas as etapas introduzidas na Seção 3.2.4. Nelas são utilizadas notações para o detalhamento do mecanismo. Essas notações são apresentadas na tabela abaixo:

Quadro 3. Notações dos Elementos Envolvidos na Descrição do Mecanismo.

NOME	DESCRIÇÃO	NOME	DESCRIÇÃO
$\boldsymbol{CSP}$	Serviço de armazenamento de dados em nuvem. Ele é o *SP* da federação.	$\boldsymbol{cf_i}$	Arquivo incluído em uma caixa de interesse. Ele será fragmentado em k partes pelo mecanismo de deduplicação para ser armazenado em nuvem ($1 \leq i \leq n,\ cf_i \in$

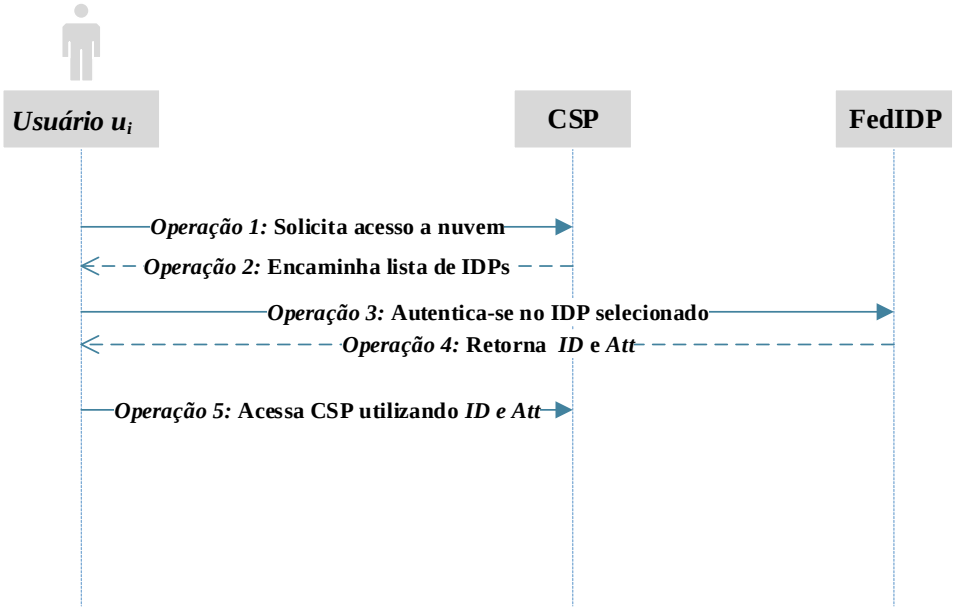
			$CIF, \lvert CIF \rvert = n)$
FEDIDP	Serviço de identidade federada. Ele disponibiliza o *IDP* da federação.	***CIcomp***	Conjunto de caixas de interesses que possuem os mesmos atributos que $c_i,\ CIcomp \subseteq CI$
CIGER	Serviço gerenciador de caixas de interesses.	CI**selec**	Conjunto de caixas de interesses selecionadas por $u_i,\ CI\text{selec} \subseteq CI\text{comp}$.
ID, ATT	Identificador de acesso e atributos do usuário.	$\boldsymbol{c_i}$	Caixa de Interesses de um usuário u_i $(1 \leq i \leq n,\ c_i \in CI, \lvert CI \rvert = n)$.
***DEDUP*()**	Serviço de deduplicação de dados.	F**comp**	Conjunto de arquivos recebidos através do compartilhamento $F\text{comp} \subseteq F$.
$\boldsymbol{u_i}$	Usuário do serviço de armazenamento de dados em nuvem $(1 \leq i \leq n,\ u_i \in U, U{=}n.$	$\boldsymbol{H(cf_i)}$	Valor hash de cf_i.

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 apresentado na Seção 2.4. A Figura 17 sintetiza esta

etapa:

Figura 16. Etapa de Autenticação e Obtenção de Atributos (Diagrama de Sequencias *UML*)



Fonte: Elaborada pelo Autor

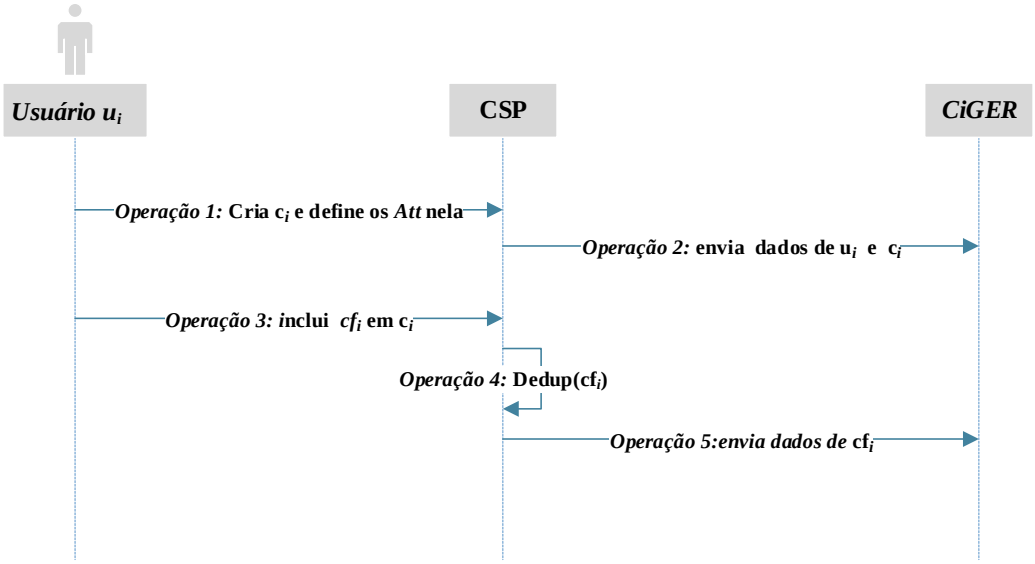
Inicialmente o usuário u_i solicita acesso ao serviço de armazenamento de dados em nuvem *CSP* (*Operação 1*). Então, a nuvem redireciona o usuário para um *IDP* da federação para que ele se autentique e obtenha acesso ao serviço (*Operação 2 e Operação 3*). Obtendo sucesso na autenticação, o usuário recebe sua credencial e seus atributos (*Operação 4*) e então comunica-se com a nuvem novamente para acessar seu espaço de armazenamento (*Operação 5*).

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 18 apresenta o funcionamento desta etapa.

Figura 17. Etapa de Criação de Caixas de Interesses (Diagrama de Sequencias *UML*)



Fonte: Elaborada pelo Autor

A etapa é iniciada pelo usuário quando ele cria uma caixa de interesses na nuvem (*Operação 1*). Para isso ele poderá tanto selecionar um diretório na nuvem e definir seus metadados, quanto adicionar os metadados de uma caixa de interesse em um diretório existente. Os metadados gerais que podem ser definidos são nome, data de criação e descrição da caixa de interesse. Os metadados específicos são os atributos da federação. Em seguida, o componente C*SP* informa o componente *CIGer* que o usuário u_i criou a caixa de interesses 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 outro diretório qualquer (*Operação 3*). O *CSP,* ao receber o arquivo, executará o serviço de deduplicação (*Operação 4*). Este serviço realizará a verificação de redundância e atualizará os metadados de rastreamento do arquivo caso necessário (ex. dados de localização de um arquivo novo). Além dessas atualizações, o deduplicador também atualizará uma tabela com os valores *hashs* dos

arquivos deduplicados. Essa tabela será utilizada na Etapa III.

Caso o arquivo não seja encontrado uma cópia do arquivo na nuvem, o serviço de deduplicação não atualiza o mapa de registros deduplicados. Isso porque o arquivo ainda não existe na nuvem e, portanto, não existem caixas de interesses com arquivos em comum para realizar o compartilhamento.

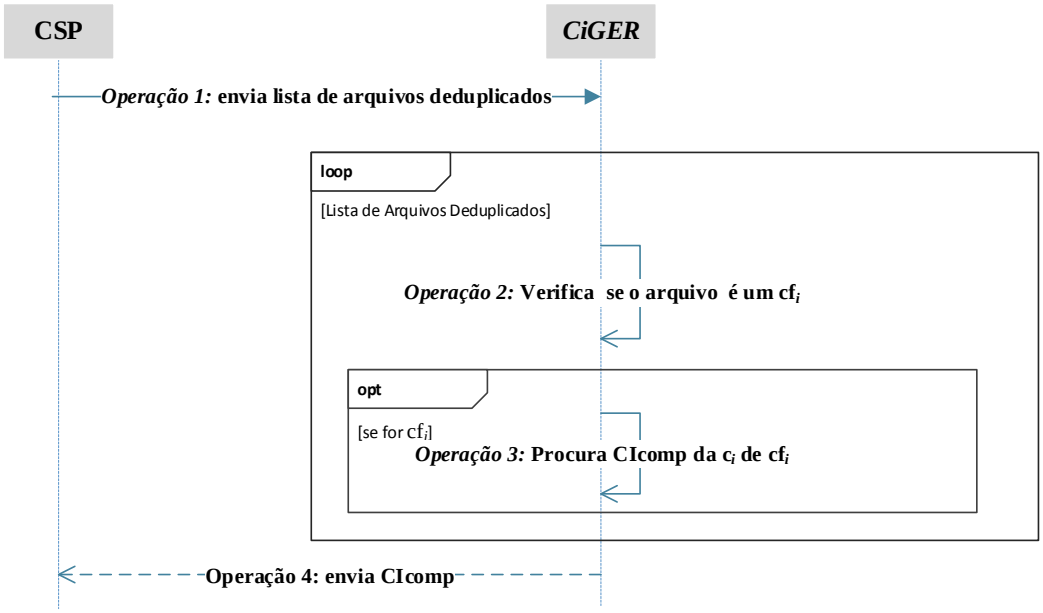
Por fim, O *CSP* termina esta etapa informando ao componente *CIGer* qual foi o arquivo enviado pelo usuário (*Operação 5*). Desta forma, este componente mantem o controle dos arquivos existentes nas caixas de interesses da nuvem registrando as relações de posse daquele usuário.

III. Etapa de Identificação de Usuários

Na etapa de identificação de usuários o *CSP* e o componente *CIGer* interagem de forma assíncrona a fim de identificar os usuários que poderão compartilhar dados entre si. Ela consiste em um processamento realizado pelo *CIGer* para a identificação de caixas de interesses que possuem arquivos em comum e na notificação da nuvem da existência dessas caixas.

A Figura 19 apresenta o funcionamento desta etapa.

Figura 18. Etapa de Identificação de Usuários (Resumo)



Fonte: Elaborada pelo Autor

A identificação usuários para o compartilhamento inicia quando a nuvem *CSP* envia a lista de arquivos dedeplucados para o componente *CIGer* (*Operação 1*). O *CIGer* extrai a lista de *hashs* desta tabela e verifica quais deles pertencem a caixas de interesses (*Operação 2*). Essa verificação é realizada comparando os *hash* que extraiu com o seu mapeamento de arquivos e caixas de interesses. Ao encontra *o hash* de um arquivo existente em uma caixa (cf_i), o componente *CIGer* obtém as caixas que também possuem o mesmo arquivo e atributos em comum (*Operação 3*).

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.

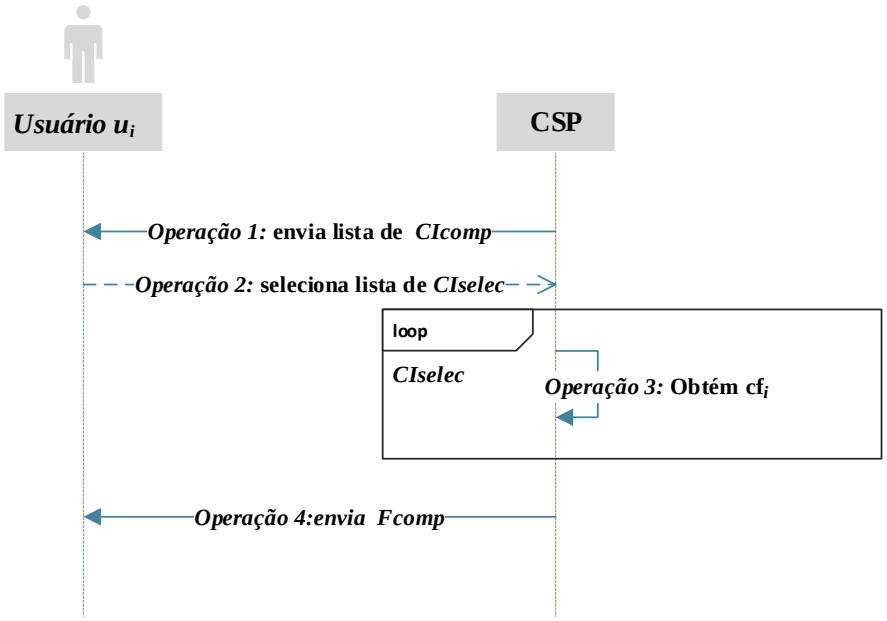
Tendo resgatado as caixas de interesses em comum, CIGer envia essa informação ao *CSP* (*Operação 4*) e, então, se inicia a Etapa 4.

IV. Compartilhamento de Dados

Esta consiste na última etapa do mecanismo. Nela o *CSP* notifica o usuário sobre a existência de caixas de interesses compatíveis com a dele (isto é, caixas de interesses com de outros usuários com arquivos em comum). Mediante o aceite, o *CSP* promove o compartilhamento de dados de uma caixa de interesse para outra.

A Figura 20 apresenta o protocolo de funcionamento desta etapa.

Figura 19. Etapa de Compartilhamento de Dados (Diagrama de Sequencias *UML*)



Fonte: Elaborada pelo Autor

Inicialmente, o serviço de nuvem realiza a *Operação 1* do fluxo acima. Esta operação refere-se à 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 aceita por u_i são encaminhadas para a nuvem (*Operação 2*).

Recebendo a resposta do usuário, o *CSP* realiza a *Operação 3*. Nela a nuvem busca os arquivos cf_i de cada caixa de interesse selecionada. As caixas de interesses que foram rejeitadas são removidas da lista de sugestões de sugestões mantida por *CSP*. Então, por fim, C*SP* disponibiliza os arquivos pertencentes as caixas de interesses aceitadas por u_i *(Operação 4)*. Essa disponibilização é realizada através da inclusão da caixa selecionada na conta de u_i.

3.3 Prova de Conceito

Tomando como base as definições e etapas descritas anteriormente neste capítulo, nesta seção apresenta-se o protótipo do mecanismo de compartilhamento por caixas de interesses.

Inicialmente, na Seção 3.3.1, são apresentadas as características de compatibilidade do protótipo com o sistema Owncloud[15]. O Owncloud consiste em um projeto *open source* no contexto de nuvens computacionais que provê formas de criação e gerenciamento de serviços de nuvens de armazenamento. O protótipo desenvolvido esta baseado em alguns aspectos desse sistema a fim de simulá-lo ao executar o mecanismo de compartilhamento por caixas de interesses.

Em seguida, na Seção 3.3.2. é apresentada a arquitetura do protótipo. Nela descreve-se como os componentes e etapas do mecanismo são realizadas por ele.

3.3.1 Compatiblidade com o Owncloud

O protótipo implementado é baseado na arquitetura de armazenamento de dados utilizada pelo sistema Owncloud. Este projeto mostrou-se interessante para este trabalho, pois oferece suporte a customização de seu serviço através de *plugins* denominados de *Owncloud Apps*. Esses *plugins* facilitam a implementação do mecanismo proposto em um cenário real.

O protótipo deste trabalho, não foi desenvolvido na forma de um *Owncloud App*. Apesar disso, ele considera características da arquitetura do Owncloud. Essas características foram introduzidas para que o protótipo, mesmo não sendo integrado ao Owncloud, possa simular a execução de um *plugin*. Desta forma, o protótipo pode ser traduzido para a linguagem de programação nativa do Owncloud a fim de ser integrado como um *Owncloud App*.

As características do Owncloud utilizadas na implementação do protótipo são

[15] https://owncloud.com/

identificadas abaixo.

Utilização do diretório files do Owncloud

O Owncloud fornece seus serviços por meio de duas formas principais de implantação. Na primeira, ele pode ser utilizado como um serviço que gerencia arquivos armazenados em uma infraestrutura de armazenamento própria do usuário. Esta forma é a ideal para a criação de nuvens privadas estabelecidas sem o auxilio de sistemas de armazenamento de dados em nuvem. Por outro lado, o Owncloud pode ser utilizado como um *middlware* que gerencia arquivos em diferentes serviços de armazenamento nuvem. Esta forma é ideal para o gerenciamento de arquivos em nuvens privadas estabelecidas com alguns sistemas de armazenamento em nuvens disponíveis no mercado (ex. Openstack Swift) e em nuvens públicas (ex. Dropbox).

Apesar das diferentes formas de implantação, o Owncloud utiliza um diretório específico para armazenar dados enviados pelos usuários. Este diretório, denominado por padrão de *files*, tem a função de representar a raiz do espaço de armazenamento de dados do Owncloud. Quando utilizado em um ambiente local, o diretório files é apenas um diretório local gerenciado pelo Owncloud. Quando utilizado com um provedor externo ele é sincronizado com uma estrutura de agrupamento de dados que esses provedores oferecem, como por exemplo, *buckets* do Amazon S3 e *containers* do Swift Openstack. Neste segundo caso, o diretório files reflete o estado da conta do usuário com que está sincronizado.

Tendo isso em vista, o protótipo desenvolvido simula a forma de armazenamento de dados realizada pelo Owncloud. O protótipo considera a forma de armazenamento de dados utilizando recursos locais montados no diretório *files*.

Organização dos Arquivos

Conforme apresentado acima, o Owncloud organiza os arquivos armazenados pelos usuários no diretório *files*. Outra característica da forma que o Owncloud armazena os dados dos usuários está na organização interna deste diretório. Ele possui subdiretórios que organizam o espaço de armazenamento por usuários do sistema. Cada subdiretório

possui o nome de um usuário do sistema e armazena os dados do respectivo usuário cujo nome é utilizado pelo diretório. Para fins de exemplificação, um arquivo arquivo.pdf de um usuário denominado usuario1 é armazenado da seguinte forma: */files/usuário1/arquivo.pdf*.

Tendo isso em vista, o protótipo desenvolvido mantém o padrão estrutural de armazenamento de arquivos do Ownlcoud ao armazenar os dados dos usuários.

Utilização do banco de dados do Owncloud

O Owncloud utiliza o banco de dados para: auxiliar no rastreamento de arquivos armazenados, gerenciar seus usuários e para armazenar parâmetros de configuração do serviço.

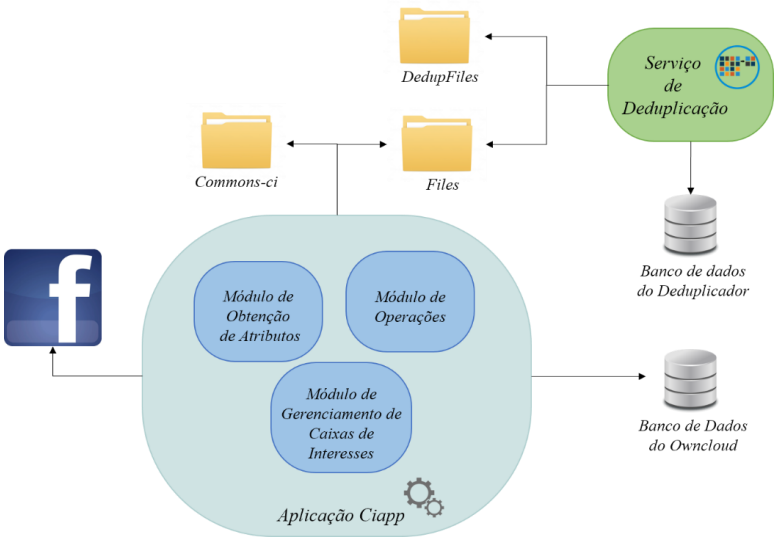
O protótipo desenvolvido, quando necessário, executa suas operações relacionadas ao serviço de nuvem (ex. upload de arquivos, criação de diretório) e atualiza informações em uma base de dados com tabelas iguais a do Owncloud. Além disso, esse banco de dados recebeu algumas modificações a fim de atender a funcionalidade de compartilhamento por caixas de interesses, como por exemplo, a criação de tabelas adicionais para o gerenciamento das caixas de interesses e sugestões de caixas compatíveis.

Mediante as considerações apresentadas, a seção seguinte apresenta uma visão geral da arquitetura do protótipo desenvolvido.

3.3.2 Visão Geral da Arquitetura do Protótipo

Na Seção 3.1 foram introduzidos os principais conceitos, os componentes envolvidos no mecanismo de compartilhamento por caixas de interesses assim como o fluxo de interações entre eles. Tendo isso em vista, O protótipo desenvolvido implementa esses elementos. A Figura 21 apresenta uma visão geral de como o protótipo realiza tais componentes e suas interações.

Figura 20. Visão Geral da Arquitetura do Protótipo



Fonte: Elaborada pelo Autor

Conforme se pode observar na Figura 21, o protótipo consiste em dois componentes principais: a aplicação Ciapp e o serviço de deduplicação.

A aplicação Ciapp é o serviço que implanta o mecanismo de compartilhamento de dados. Neste protótipo, ela simula o sistema Owncloud. Essa aplicação é composta por três subcomponentes: o módulo de operações, o módulo de obtenção de atributos e o módulo de gerenciamento de caixas de interesses.

O componente de deduplicação de dados consiste em um *daemon* que monitora o diretório *files*. Conforme apresentado inicialmente, este é o diretório em que a aplicação armazena os dados dos usuários. O Ciapp atualiza esse diretório com novos arquivos e o deduplicador os move para a pasta *DedupFiles*. Apenas uma cópia de cada arquivo é armazenada neste último diretório. Adicionalmente, este componente cria arquivos no diretório *files* que mapeiam o relacionamento do usuário como seus dados que foram armazenados ali. Desta forma a aplicação Ciapp necessita apenas ler esses arquivos para disponibilizar a listagem de arquivos para seus usuários.

Sempre que um arquivo redundante é identificado pelo componente deduplicador ele atualiza o mapeamento no diretório *files*, remove o arquivo e registra seus metadados no seu banco de dados. Esse metadados possuem informações do arquivo assim como a sua localização no diretório *DedupFiles*. Por fim, o deduplicador também atualiza uma tabela com a lista de arquivos deduplicados. Essa tabela armazena os

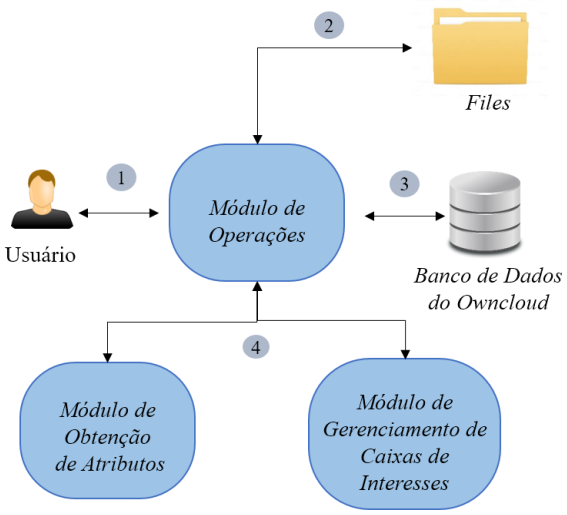
dados que serão informados a aplicação Ciapp quando estiver executando a Etapa III do mecanismo. Ela identifica cada arquivo deduplicado pelo seu valor *hash*.

Tendo em vista que os macro componentes da aplicação Ciapp e o serviço deduplicador foram apresentados, a seguir, são apresentados os módulos que compõe o Ciapp.

3.3.2.1 O Módulo de Operações

Este módulo realiza as tarefas de um serviço de armazenamento de dados em nuvem. Através dele é possível enviar, remover e listar arquivos, por exemplo. Além disso, através dele a aplicação Ciapp disponibiliza os recursos de criação de caixas de interesses, de definição de atributos nas caixas de interesses e de aceitação ou rejeição de um arquivo sugerido pela nuvem. Para tal, este módulo manipula o diretório *files* e o banco de dados do Owncloud para refletirem as operações dos usuários. A Figura 22 apresenta uma visão geral de Funcionamento deste módulo.

Figura 21. Visão Geral de Funcionamento do Módulo de Operações



Fonte: Elaborada pelo Autor

Conforme ilustrado, o módulo de operações é o responsável por receber as requisições dos usuários (item 1). Após recebê-las, ele poderá realizar quatro tipos de operações diferentes, que são: operar sobre o diretório *files* (item 2), operar sobre o

banco de dados do *Owncloud* (item 3) ou interagir com o módulo de obtenção de atributos e interagir com o módulo de gerenciamento de caixas de interesses (item 4).

As operações sobre o diretório *files* e sobre o banco de dados do Owncloud ocorrem mediante as solicitações relacionadas ao gerenciamento de arquivos e ao compartilhamento de dados.

Na operação de envio de dados, o módulo armazena o arquivo no subdiretório de *files* correspondente ao usuário conforme o padrão /*files*/<*usuario*>/<*caminho__do_arquivo*> do Ownlcoud. Após essa operação o deduplicador verifica se existe redundância de dados conforme descrito anteriormente.

Na operação de remoção, o módulo apenas notifica o deduplicador para que este atualize seus metadados, e em seguida remove o mapeamento existente no diretório *files*. Junto a essas operações, o módulo atualiza a base de dados da aplicação. Esta atualização ocorre conforme o padrão do *Owncloud.*

Na operação de listagem de dados, o módulo lê os dados mapeados no diretório *files*. Cada usuário possui seu mapeamento, e, portanto, o módulo apenas carrega o arquivo de mapeamento do usuário solicitante.

Na criação de caixas de interesses o módulo de operações cria um diretório dentro do correspondente subdiretório de *files* do usuário. Ele também registra os metadados gerais e específicos da caixa de interesses no banco de dados do Owncloud. Para listagem de caixas de interesses, a aplicação carrega os diretórios existentes para aquele usuário e verifica quais foram registrados como caixas de interesses pelos seus metadados no banco de dados.

Por fim, através desse módulo é realizada a aceitação e a rejeição do compartilhamento de caixas de interesses. Essa operação é iniciada com a leitura da tabela de sugestões de compartilhamento mantida na base de dados da aplicação. As sugestões lidas são disponibilizadas ao usuário que pode aceitá-las ou rejeitá-las. Ao aceitar, a nuvem verifica na base de dados os arquivos existentes na caixa de interesses do emissor, e cria referências deles para o receptor do compartilhamento. Ao rejeitar, a aplicação apenas remove a sugestão da base de dados.

Além dessas operações, o módulo de operações atua como intermediário entre o usuário e os demais módulos. Por exemplo, ele é responsável por receber a solicitação de obtenção de atributos realizada pelo usuário e de solicitar essa operação ao módulo

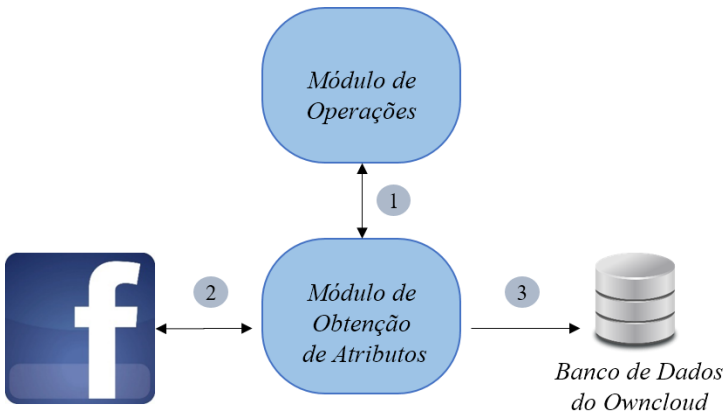
de obtenção de atributos. Ele também é responsável por fornecer dados das caixas de interesses do usuário para o módulo de gerenciamento de caixas de interesses quando o usuário adiciona arquivos nelas.

A seguir são apresentados os módulos de obtenção de atributos e o de gerenciamento de caixas de interesses, respectivamente.

3.3.2.2 O Módulo de Obtenção de Atributos

Este módulo é responsável por se comunicar com um serviço de federação e obter os atributos dos usuários. No protótipo, o módulo usufrui dos serviços da rede social Facebook a fim de obter esses atributos. Isto é possível, pois esta rede social prover seu serviço de autenticação e autorização por meio do protocolo OAuth. A Figura 23 apresenta uma visão geral do módulo de obtenção de atributos.

Figura 22. Visão Geral de Funcionamento do Módulo de Obtenção de Atributos



Fonte: Elaborada pelo Autor

Em conformidade com o que foi apresentado na Etapa I do mecanismo (ver Seção 3.2), a aplicação Ciapp interage com o Facebook para obter os atributos dos usuários. O módulo de obtenção de atributos recebe a solicitação a partir do módulo de operações (item 1). Ele realiza as requições ao Facebook para obter os atributos do usuário (item 2). Após obtê-los, o módulo de obtenção de atributos armazena-os na base de dados do Owncloud (item 3). A partir de então esses atributos são disponibilizados para o usuário

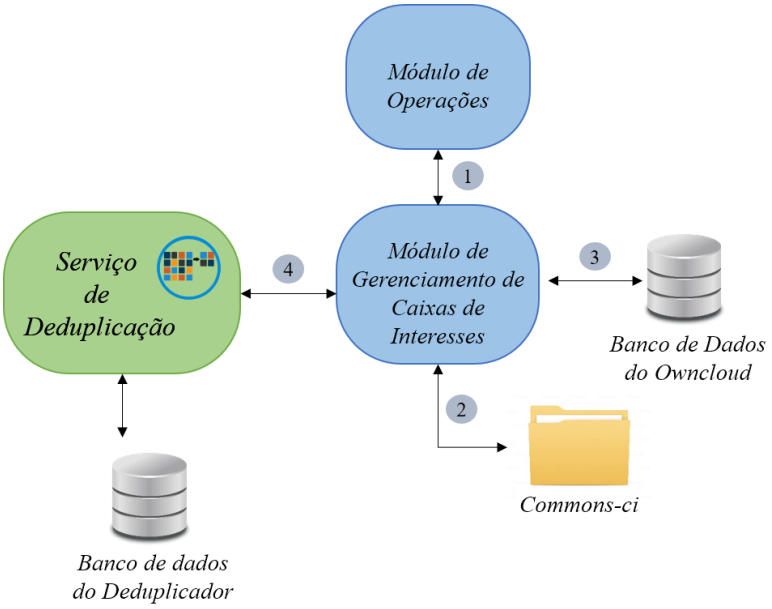
sempre que ele criar uma caixa de interesses.

3.3.2.3 O Módulo de Gerenciamento de Caixas de Interesses

O módulo de gerenciamento de caixas de interesses é o componente *CiGer* apresentado na Seção 3.2. Ele atua de duas formas no mecanismo: monitorando as operações dos usuários referentes à criação e modificação de caixas de interesses e procurando por usuários que possuem arquivos iguais em caixas de interesses com atributos em comum.

A Figura 24 apresenta uma visão geral do módulo de gerenciamento de caixas de interesses.

Figura 23. Visão Geral de Funcionamento do Módulo de Gerenciamento de Caixas de Interesses



Fonte: Elaborada pelo Autor

A Figura 24 apresenta as interações do módulo de gerenciamento de caixas de interesses em suas duas formas de atuação. Na primeira, ele recebe informações vindas do módulo de operações toda vez que um usuário modifica sua caixa de interesses, seja adicionando atributos, seja adicionando ou removendo arquivos (item 1). Essas informações são utilizadas pelo módulo de gerenciamento de Caixas de interesses para

produzir o mapeamento de quais usuários apresentam caixas de interesses com arquivos em comum. Elas são registradas no diretório *commons-ci* em arquivos denominados *commons-cis.map* (item 2).

Na segunda forma de atuação, o módulo de gerenciamento de caixas de interesses atua como um serviço assíncrono que verifica a lista de arquivos deduplicados criada pelo componente deduplicador. Ele solicita ao serviço de deduplicação essa lista de arquivos deduplicados. Essa verificação consiste na leitura dos *hash* de cada arquivo deduplicado em intervalos de tempo constantes (ex. a cada 12h, uma vez ao dia).

Após obter a lista de *hashs*, o módulo verifica quais deles correspondem a arquivos armazenandos em caixas de interesses. A verificação ocorre comparando os *hash* com os dados mapeados no diretório *commons-ci*. Por fim, os usuários que possuem dados iguais armazenados em caixas de interesses com atributos em comum são registrados na lista de sugestões de compartilhamento por caixas de interesses na base de dados da aplicação. Essa lista de sugestões é utilizada pelo módulo de operações para informar os usuários sobre outros usuários que suscetíveis ao compartilhamento por caixa de interesses.

4. CONSIDERAÇÕES SOBRE O MECANISMO

Tendo em vista o mecanismo de compartilhamento por caixas de interesses introduzido no Capítulo 3, este Capítulo apresenta considerações relativas aos cenários de aplicação e ao desempenho desse mecanismo.

Inicialmente, na Seção 4.1 são apresentadas discussões relativas aos possíveis cenários de aplicação do mecanismo. Dois cenários são apresentados: o compartilhamento de dados pessoais entre indivíduos e a nuvem como serviço de fornecimento de conteúdo.

Na Seção seguinte são apresentados discussões e resultados obtidos nos testes realizados com o protótipo apresentado na Seção 3.3. Os aspectos discutidos são o impacto do mecanismo de compartilhamento por caixas de interesses na deduplicação de dados e os fatores que influenciam na etapa de identificação de usuários com interesses em comum.

4.1 Cenários de Aplicação

Nesta seção identificam-se diferentes cenários em que o mecanismo proposto pode ser aplicado. Conforme descrito na Seção 3, o compartilhamento por caixas de interesses permite que os usuários que não se conhecem possam compartilhar dados desde que apresentem potenciais interesses em comum. Este é o cenário básico de aplicação. Além deste, outra aplicação é apresentada a fim de destacar os benefícios que o mecanismo pode oferecer por meio das nuvens computacionais.

Os cenários serão apresentados nas próximas seções. Cada seção será introduzida com uma descrição de uma situação envolvendo usuários dentro de um cenário

proposto. Essas descrições tem a finalidade de facilitar o entendimento da aplicação. Em seguida, após cada descrição, destacam-se os benefícios dos cenários propostos.

4.1.1 Compartilhamento de Dados Pessoais entre Indivíduos

O cenário

Duas pessoas, João e Maria, 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. Maria é uma usuária que armazena, na nuvem, documentos referentes à sua pesquisa. Embora alguns documentos de Maria 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).

João, por sua vez, é um pesquisador da mesma área de conhecimento que Maria, mas que não a conhece. Nesse contexto, Maria e João são usuários que podem se beneficiar através do compartilhamento de dados por caixas de interesses. Para tal, basta que ambos criem uma caixa de Interesses com o atributo pesquisador[16] no serviço de armazenamento de dados em nuvem. Então, ao adicionar seus materiais de pesquisa (ex. apresentações, artigos próprios e artigos de referencia) nestas caixas a nuvem de armazenamento verificará se existem arquivos em comum entre eles. Caso exista, a nuvem fará sugestões para ambos os usuários para que realizem o compartilhamento de arquivos entre si.

Através do cenário apresentado é possível destacar o principal benefício deste trabalho. João e Maria são potenciais usuários que podem se beneficiar do compartilhamento de dados em nuvem. Isso porque são pessoas que possuam interesses

[16] Considera-se que todo o pesquisador nesse cenário possui o atributo “pesquisador”.

profissionais na mesma área de pesquisa. Por outro lado, eles não se conhecem. Esse fator dificulta o compartilhamento de arquivos entre eles.

O compartilhamento por caixas de interesses trata essa limitação. Conforme observado, para que ambos os usuários compartilhem dados através desse mecanismo basta que eles criem "pastas" (conforme apresentado na Seção 3.2, as caixas de interesses são pastas com metadados diferenciados) e definam atributos que concretizem sua intenção de compartilhamento. Por exemplo, ao desejar trocar arquivos com familiares, o atributo definido poderia ser "família", ao desejar trocar com pesquisadores de um mesmo departamento o atributo poderia ser "pesquisador-depx".

Por outra perspectiva, os atributos fornecem o controle do usuário sobre seus arquivos também. Se Maria ou João removerem o atributo de sua caixa de interesses (ou trocarem), ela se torna indisponível para o compartilhamento, ou torna-se disponível para outro contexto de compartilhamento (ex. se o atributo novo for pesquisadores-biologia-ufpa o compartilhamento estaria restrito a esse contexto).

4.1.2 *Nuvens de Armazenamento como Serviços de Fornecimento de Conteúdo*

O cenário

A PubCloud consiste em uma nuvem pública em que diversos tipos de usuário armazenam dados para diferentes propósitos. João é usuário desse serviço de armazenamento de dados em nuvem. Ele a utiliza para armazenar dados pessoais através de seu computador e de seu *smartphone*. Outro usuário que usufrui dos serviços de PubCloud é a empresa Acme. Embora esta empresa possua documentos que devam ser mantidos de forma privada (ex.registros financeiros) outros são destinados aos seus usuários (ex. manuais de seus produtos, revistas de divulgação). Acme, então, utiliza os serviços de PubCloud para armazenar os seus arquivos destinados aos usuários.

Nesse contexto, João e a empresa Acme são usuários que podem se beneficiar através do compartilhamento de dados por caixas de interesses. Acme pode criar uma

ou mais caixas de interesses para compartilhar os dados mantidos na nuvem pública com seus usuários. Acme então deve definir atributos nela(s) que seus usuários também possam definir na sua caixa de interesses. Para tal o serviço de nuvem deve possuir integração com um provedor de atributos público, como por exemplo, uma rede social. Acme então pode definir como atributo o nome de sua organização nesta rede social. João por sua vez, define o mesmo atributo na sua caixa de interesses. A nuvem, então, sugerirá a João a caixa de interesses da empresa Acme. Deste modo, Acme poderá disponibilizar dados ao seu usuário de forma proativa.

O cenário apresenta uma forma de como organizações podem usufruir da popularidade dos serviços públicos de armazenamento de dados em nuvem por meio do mecanismo de compartilhamento por caixa de interesses. Através deles, elas podem fornecer conteúdo aos seus usuários de forma proativa, evitando, por exemplo, que eles visitem o site de sua companhia ou outro canal de comunicação para obter tais conteúdos. O usuário, por sua vez, tem total controle sobre aceitar ou não os dados das organizações. Esses arquivos podem ser obtidos de acordo com suas necessidades.

Por outro lado, a utilização de nuvens públicas tem apresentado resistência de utilização por parte das organizações. Isso devido a possível quebra de privacidade que podem ocorrer sobre dados armazenados na sua infraestrutura. Apesar disso, a aplicação apresentada acima prevê o armazenamento de arquivos que não apresentam restrições de privacidade por parte da organização. De fato, os dados que são incentivados a serem compartilhados são os conteúdos públicos da empresa. Deste modo, mesmo que a nuvem não possua um serviço de armazenamento com políticas e mecanismo rígidos de segurança, isso não apresenta um risco de segurança para as organizações.

4.2 Discussões Sobre o Desempenho do Mecanismo

Esta seção apresenta considerações sobre o mecanismo de compartilhamento por caixas de interesses a partir de testes realizados com o protótipo desenvolvido. Esses testes foram realizados a partir um computador com as seguintes configurações:

Quadro 4. Configurações do Ambiente de Teste do Protótipo.

CONFIGURAÇÕES	
SISTEMA OPERACIONAL	*Windows 8.1*
QUANTIDADE DE MÉMORIA	*6 GB*
CPU	*Core I7 Quatro Núcleos de 2.4 GHz de frequência*

Os resultados dos testes foram obtidos com base na média de 10 execuções de casa caso. Esses 10 resultados também foram utilizados para o calculo do desvio padrão. O desvio padrão está representado nos gráficos por um eixo com limites superior e inferior na parte superior de cada pilar. Esse limites representam, respectivamente, o desvios padrão para mais e para menos de cada resultado.

4.2.1 Impacto do Mecanismo de Compartilhamento por Caixas de Interesses na Deduplicação de Dados

Conforme apresentado na Seção 2.3, a deduplicação de dados é um mecanismo voltado para o aprimoramento no armazenamento de dados. Para isso, o serviço de deduplicação deve constantemente monitorar os dados armazenados nos servidores e aplicar as devidas operações de remoção de redundância e atualização de metadados de acordo com a necessidade.

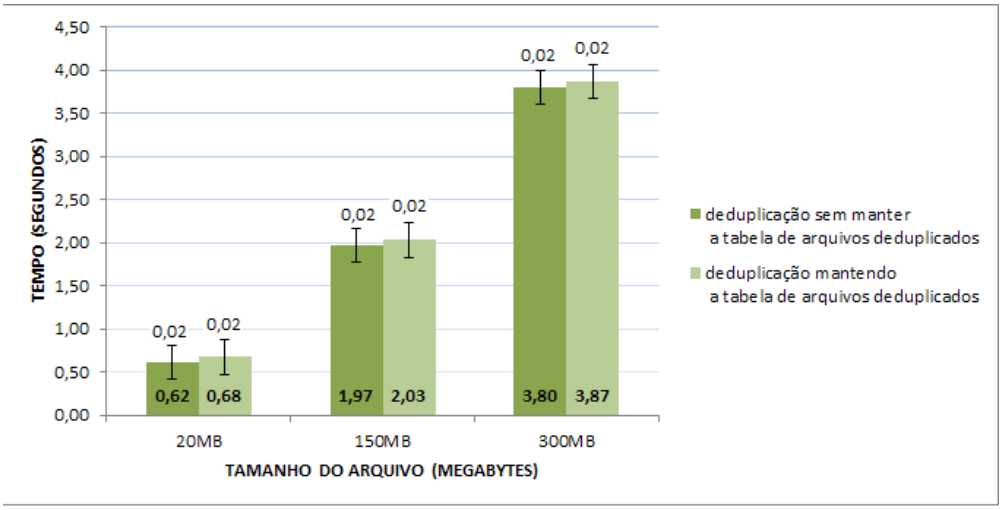
Tendo em vista que o mecanismo de compartilhamento por caixas de interesses usufrui da deduplicação para obter suporte ao seu funcionamento, verificou-se, então, o impacto com relação ao tempo que ele causa na execução do serviço de deduplicação. Este é um fator importante para a viabilidade da proposta. Isto porque caso haja um grande acréscimo no tempo de execução da deduplicação, utilizar tal estratégia se torna inviável. Um exemplo em que o tempo de deduplicação é um fator indispensável, está

no caso de deduplicação na fonte em quem o processo de envio de arquivos do usuário depende do termino da execução da deduplicação.

Para o referido teste considerarou-se o seguinte cenário: três arquivos com tamanhos distintos (20, 150 e 300 Megabytes) foram submetidos a execução do serviço de deduplicação produzido durante esta pesquisa. A primeira execução calculou o tempo de deduplicação de dados sem a presença da alteração no serviço voltada para o auxílio ao compartilhamento por caixas de interesses. Em outras palavras, na primeira execução, o deduplicador não mantinha a lista de dados deduplicados. Já na segunda execução o tempo foi calculado com esta modificação.

Nesse contexto, o Gráfico 1 apresenta o resultado obtido nos testes através do protótipo.

Gráfico 1. Resultado do Teste de Impacto do mecanismo de compartilhamento por Caixas de Interesses sobre a deduplicação de Dados.



Conforme se pode observar pelos resultados obtidos, a adição da lista de dados deduplicados não gerou impacto representativo sobre o tempo de execução serviço de deduplicação de dados. Em todos os caso testados a deduplicação torna-se 0,06 segundos mais lenta para auxiliar o compartilhamento por caixas de interesses (com desvio padrão de 0,02 segundos). Essa característica mostrou-se não ser influenciada pelo tamanho do arquivo. Isso se explica por que o processo de manutenção da lista de arquivos deduplicados consiste apenas em uma instrução de armazenamento do *hash* do arquivo na base de dados do deduplicador juntamente com a data de sua inserção. Como

este serviço produz o *hash* durante o processo de deduplicação (tarefa mais custosa computacionalmente), então, este valor é aproveitado para esta atualização.

4.2.2 Fatores que Influenciam na Etapa de Identificação de Usuários com Interesses em comum

Conforme descrito na Seção 3, a identificação de usuários com interesses em comum corresponde a etapa em que o serviço de nuvem produz as sugestões de compartilhamento de caixas de interesses entre os usuários. Nesta etapa, o gerenciador de caixas de interesses deve procurar por arquivos deduplicados e em seguida verificar quais usuários possuem interesses em comum através de seu mapeamento.

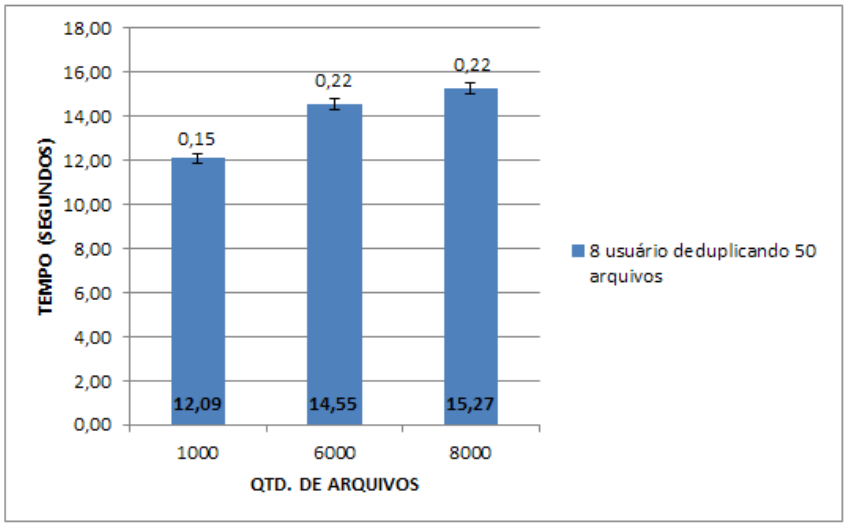
Nesta subseção, apresentam-se os resultados obtidos nos testes de impacto no tempo de execução dessa etapa mediante o aumento na quantidade de usuário e de aumento da quantidade de arquivos na nuvem. Esses elementos influenciam diretamente o funcionamento da etapa de identificação de usuários, pois o CiGer possui o mapeamento de quais usuários possuem arquivos em comum.

No primeiro teste realizado, buscou-se verificar o quanto a quantidade de arquivos não deduplicados da nuvem influenciava no tempo de execução da etapa de identificação de usuários. Para tal, simularam-se três ambientes em que a nuvem possuía 1000, 6000 e 8000 arquivos não redundantes entre os usuários. Nestes ambientes foram criados 8 usuários com 50 arquivos repetidos em caixas de interesses e que, portanto, geram sugestões de compartilhamento.

Por meio deste teste buscou-se evidências sobre o comportamento do mecanismo caso houvessem muitos arquivos, porém poucos usuários com arquivos em comum em caixas de interesses. Este corresponde ao cenário em que poucos usuários utilizam o compartilhamento por caixas de interesses no serviço de armazenamento de dados em nuvem.

O Gráfico 2 apresenta os resultados obtidos através do teste realizado:

Gráfico 2. Impacto do Aumento de Arquivos não redudantes na Etapa de Identificação de Usuário.



Conforme se pode observar no Gráfico2, o resultado dos testes, apresentam que a quantidade de arquivos que são redundantes na nuvem produz um pequeno impacto na etapa de identificação de usuário. A diferença de tempo na execução desta etapa em um cenário de 1000 arquivos, para outro de 8000 foi de pouco mais de 3s.

Esse resultado justifica-se pelo processamento que ocorre no momento da produção de sugestões. A quantidade de arquivos influencia apenas na busca das caixas de interesses dos usuários emissores do compartilhamento no banco de dados da aplicação. Esta busca consiste em uma pesquisa na base de dados da aplicação. Neste cenário quanto maior a quantidade de registros armazenados nesta base de dados, maior será o tempo de execução. Em contrapartida, o tempo de execução é otimizado pelo próprio banco de dados.

Os resultados sugerem, portanto, que em um cenário em que existe uma nuvem de armazenamento em que não existem muitos arquivos redundantes em caixas de interesses, os usuários que usufruem do mecanismo possuirão o tempo de execução de seu compartilhamento pouco degradado pelo crescimento da nuvem.

Os testes seguintes relacionam-se com o aumento de usuários das caixas de interesses e o aumento de arquivos deduplicados dentro de caixas de interesses. Este cenário corresponde ao inverso do anterior. Isto é, nele considera-se um serviço de armazenamento em nuvem em que os usuários utilizam o mecanismo de compartilhamento por caixas de interesses com frequência.

Para os referidos testes considerou-se que a nuvem possuía um total de 1000 arquivos não redundantes armazenados. Conforme apresentado no Gráfico 2 a quantidade desse tipo de arquivo impacta pouco no tempo de execução do mecanismo, portanto, este parâmetro não foi variado durante os testes.

Além disso, na verificação de impacto do aumento de arquivos deduplicados em caixas de interesses considerou-se um cenário com uma quantidade de usuários compartilhando dados entre si invariante. Neste teste apenas 2 usuários foram utilizados. Assim, é possível verificar apenas o aumento de tempo ocasionado pela quantidade de arquivos deduplicados. Na verificação de impacto do aumento de usuários das caixas de interesses considerou-se o inverso. O número de usuário variou de 2 até 8, e os arquivos deduplicados foram mantidos constantes (50 arquivos em cada teste).

Os Gráfico 3 e 4 apresentam os resultados obtidos no teste de aumento na quantidade de usuários e na quantidade de arquivos deduplicados pertencentes a caixas de interesses.

Gráfico 3. Impacto do Aumento de Arquivos Deduplicados pertencentes a Caixas de Interesses na Etapa de Identificação de Usuário.

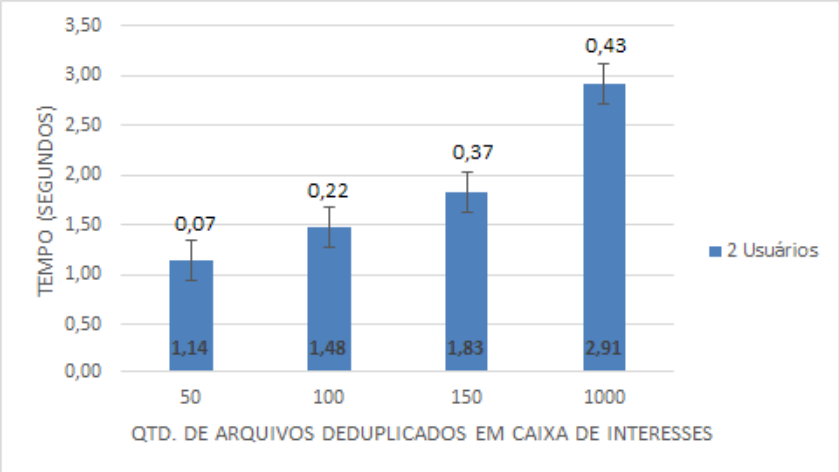
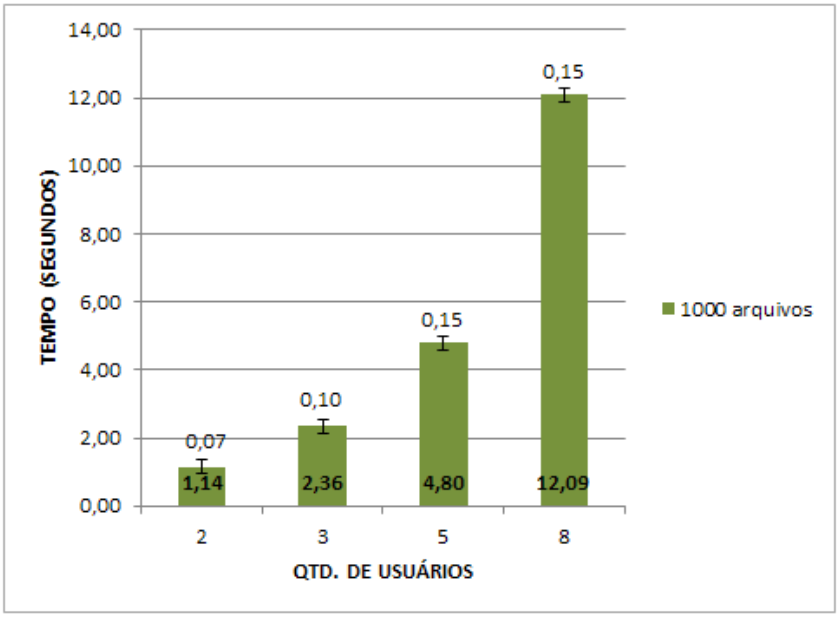


Gráfico 4. Impacto do Aumento de Usuários na Etapa de Identificação de Usuário.



Conforme os gráficos acima apresentam, em ambos os cenários houve o aumento do tempo de execução da etapa de identificação de usuários. Isso ocorre, pois o componente CiGer possui um mapeamento de todos os arquivos existentes em caixas de interesses do usuários da nuvem.

Além disso, é possível observar que o aumento da quantidade de usuários causa maior impacto no tempo de execução da terceira etapa do mecanismo do que o aumento de arquivos deduplicados. Enquanto no primeiro caso a variação de 50 para 1000 arquivos gerou um acréscimo de aproximadamente 2s, no segundo o aumento de 2 para 8 usuários gerou um aumento de aproximadamente 11s.

Este fato é explicado pela forma que o CiGer implementado no protótipo realiza o mapeamento mencionado. Ele é realizado na forma *arquivo → usuários*. Em outras palavras, para cada arquivo deduplicado o CiGer irá verificar quais usuários os possuem. Se um usuário possui mais de um arquivo deduplicado ele é descoberto, apenas, no momento que este arquivo é verificado e gera uma operação extra no banco de dados para arquivo em que for encontrado.

Por fim, conforme se observa, o compartilhamento por caixas de interesses é um mecanismo que exige um tempo de processamento maior que os demais mecanismos presentes nos serviços de nuvem. Esse comportamento era esperado. Devido a essa

característica ele apresenta a fase de identificação de usuário de forma assíncrona as interações dos usuários.

A execução da etapa de identificação de usuário de forma assíncrona oferece dois benefícios principais ao mecanismo: o primeiro é que desta forma é possível usufruir de momentos de baixo consumo de recursos computacionais para executar a etapa de identificação de usuários. Isto é, o serviço de nuvem pode definir os períodos de baixa demanda dos usuários para realizar o processamento de identificação de usuários, por exemplo.

O segundo refere-se à possibilidade de implementação do módulo de identificação de usuários como um serviço isolado da aplicação de nuvem e até do servidor em que ocorre a deduplicação de dados. Desta forma, ele pode ser incluído em um servidor dedicado com *hardware* específico que forneça suporta para que execute suas operações mais rapidamente.

Desta forma, o compartilhamento por caixas de interesses pode atuar de forma complementar as estratégias tradicionais de compartilhamento e oferecer o recurso e compartilhamento de dados em um cenário de aplicação diferenciado.

5. CONCLUSÕES

Neste Capítulo são resumidas as principais conclusões obtidas com o desenvolvimento deste trabalho. Inicialmente, apresentam-se as considerações finais relacionadas à proposta explicada ao longo deste documento. Em seguida, apresentam-se algumas limitações identificadas no mecanismo. Por fim, sugerem-se orientações para trabalhos futuros que venham a ser elaboradas a partir do exposto por esta pesquisa.

5.1 Considerações Finais

A computação em nuvem torna-se cada vez mais forte dentro do mundo cientifico. A abrangência deste tema compreende diversas áreas da computação. O compartilhamento de dados em nuvens, em específico, é um tópico importante neste cenário e vem sendo objeto de estudo de muitos pesquisadores. Este trabalho se insere nesse contexto como uma proposta que objetiva agregar contribuições e incentivar novas discussões na área.

Nesta dissertação propôs-se um novo mecanismo para o compartilhamento de arquivos em serviços de armazenamento de dados em nuvem. O compartilhamento por caixas de interesses usufrui de federações de identidade e da deduplicação em nuvem e fornece a capacidade de compartilhamento controlado de arquivos entre usuários que não se conhecem em serviços de armazenamento de dados em nuvem. 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.

Através deste trabalho, também se apresentou um protótipo como prova de conceito do mecanismo proposto. O desenvolvimento de tal ferramenta evidencia a possibilidade de concretização do mecanismo. Além disso, através dos testes iniciais realizados sobre ele, é possível se obter uma referência dos principais fatores que influenciam seu desempenho. Por fim, apresentaram-se cenários de aplicação que fornecem uma visão abrangente das possibilidades de implantação do mecanismo proposto.

Conforme apresentado no início deste documento, esta pesquisa foi fundamentada em seis objetivos específicos que detalhavam os passos para a obtenção do objetivo geral apresentado na Seção 1.2. Estes objetivos foram todos alcançados na medida em que:

1. Foram apresentados trabalhos relacionados a esta pesquisa e seus mecanismos de compartilhamento de dados foram identificados e apresentados (ver Capítulo 1);
2. O mecanismo de compartilhamento por caixas de interesses foi elaborado e descrito (ver Capítulo 3);
3. Um protótipo do mecanismo elaborado foi desenvolvido e apresentado como prova de conceito (ver Capítulo 3);
4. Uma discussão dos principais aspectos de desempenho do mecanismo foi apresentada (ver Capítulo 4);
5. Cenários de aplicação do mecanismo proposto foram apresentados (ver Capítulo 4).

5.2 Publicações

Além de ter alcançado seus objetivos e de ter contribuído com uma nova proposta prototipada e discutida na perspectiva de seu desempenho, esta pesquisa também produziu uma publicação no XI - Simpósio Brasileiro de Sistemas Colaborativos - 2014. O artigo intitulado Caixas de interesses: um novo mecanismo para a Colaboração através de Nuvem de Armazenamento de Dados (SILVA *et. al*., 2014) confirma o interesse pelo tema na área da colaboração em computação em nuvem.

5.3 Dificuldades Encontradas e Limitações

Apesar do esforço empenhado na elaboração desta pesquisa, é certo que muito ainda pode ser realizado para seu refinamento. Devido à falta de recursos e a necessidade de conclusão da proposta, alguns aspectos não puderam ser investigados ou desenvolvidos. Um exemplo, é o fato dos testes de desempenho terem sido realizados em um ambiente com recursos computacionais limitados. Caso o cenário utilizado apresentasse configurações mais próximas dos serviços de nuvem, poderiam ser obtidas conclusões baseadas em um cenário mais realista. O cenário atual permitiu apenas a execução de teste básicos que contribuem fornecendo dados estatísticos sobre o mecanismo, mas que incentivam o prosseguimento desta pesquisa de forma mais aprofundada.

Outra limitação desta pesquisa está relacionada a área de segurança da informação. Diversos aspectos podem ser explorados no contexto da privacidade das informações e da sua integração com a deduplicação de dados. Entretanto, a fim de manter o escopo de execução desta pesquisa alinhado ao seu cronograma, não foram incluídas discussões e avaliações quanto a segurança do mecanismo. A ausência de um aprofundamento da proposta com relação a este tema incentiva novos trabalhos nesse sentido no contexto de compartilhamento por caixas de interesses. Esse tema é destacado como uma continuação desejável visto a relevância da segurança de informação no contexto de nuvens computacionais.

5.4 Trabalhos Futuros

Espera-se que esta dissertação oriente novos trabalhos a respeito do mesmo tema. Tendo isso em vista, nesta seção são apresentadas sugestões de prosseguimentos desta pesquisa. Nela são indicadas possíveis evoluções que aprofundem aspectos da colaboração entre usuários e da segurança do mecanismo.

5.4.1 Refinamento do Controle de Acesso

Um dos principais aspectos do mecanismo proposto neste trabalho está relacionado com a possibilidade do usuário poder gerenciar o acesso de terceiros aos seus arquivos, mesmo sem conhecê-los. Este aspecto provê maior confiança na utilização do mecanismo, visto que permite o usuário controlar a forma que compartilha seus dados. Tendo em vista a importância do controle de acesso no contexto desta pesquisa, seria interessante um estudo sobre possíveis aprimoramentos da forma que o usuário controla seus dados. Uma possível evolução poderia ser a utilização de níveis de confiança ao atual controle de acesso baseado em atributos. Tal abordagem permitiria que o usuário utilizasse outro recurso para definir com que poderá promover o compartilhamento de dados e aprimoraria a segurança em seu compartilhamento de arquivos.

5.4.2 Adaptações do Mecanismo para Diferentes métodos de implementação de deduplicação de dados

Visto os diferentes métodos de implementação da técnica de deduplicação de dados, seria interessante uma evolução do mecanismo que atendesse a essas diferentes abordagens. O cenário de duplicação na fonte, especificamente, seria uma evolução desejável visto sua adoção por serviços de armazenamento de dados em nuvem como por exemplo o Dropbox (2015) motivado pelo benefício relacionado a utilização de largura de banda de forma eficiente.

5.4.3 Integração com Técnicas de Deduplicação Criptografadas

A segurança da informação tem se tornado um dos principais desafios no contexto das nuvens de armazenamento. Tendo em vista sua importância seria uma evolução interessante a pesquisa sobre formas de integração do mecanismo proposto com técnicas

criptográficas como a criptografia convergente e técnicas relacionadas que proporcionassem uma camada a mais de privacidade aos usuários dos serviços de armazenamento de dados em nuvem.

REFERÊNCIAS BIBLIOGRÁFICAS

AAS - Australian Access Federation, 2014. Disponivel em: <http://aaf.edu.au/>. Acesso em: 7 Julho 2014.

AMAZON Simple Storage Service, 2012. Disponivel em: <http://aws.amazon.com/pt/s3/>. Acesso em: 10 Novembro 2014.

ARMBRUST, M. et al. A view of cloud computing. **Communications of the ACM**, 4 Abril 2010. 50-58.

AUSANKA-CRUES, R. Methods for access control: advances and limitations. **Harvey Mudd College**, 2004. Disponivel em: <https://www.cs.hmc.edu/~mike/public_html/courses/security/s06/projects/ryan.pdf>. Acesso em: 10 Agosto 2014.

AWS | Amazon Simple Storage Service (S3), 2015. Disponivel em: <http://aws.amazon.com/s3/>. Acesso em: 10 Agosto 2014.

BO, C.; LI, Z. . C. W. Research on Chunking Algorithms of Data De-duplication. **Proceedings of the 2012 International Conference on Communication, Electronics and Automation Engineering**, 2013.

BORGMANN, E. A. **On the Security of Cloud Storage Services**. Fraunhofer Institute for Secure. [S.l.]. 2012.

BOX INC. Box.com, 2015. Disponivel em: <https://www.box.com/>. Acesso em: 10 Julho 2015.

CAFE - Comunidade Academica Federada, 2014. Disponivel em: <https://portal.rnp.br/web/servicos/cafe>. Acesso em: 7 Julho 2014.

CANARIE - Canadia Access Federation, 2014. Disponivel em: <http://www.canarie.ca/en/caf/join>. Acesso em: 7 Julho 2014.

CAS Federation, 2014. Disponivel em: <http://www.uky.edu/ukit/iog/cas>. Acesso em: 9 Julho 2014.

CATON, S. et al. Foundations of Trust: Contextualising Trust in Social Clouds. **Cloud and Green Computing (CGC), 2012 Second International Conference on**, Xiangtan, Novembro 2012. 424 - 429.

CENTURY Link, 2012. Disponivel em: <http://www.centurylink.com/business/>. Acesso em: 10 Fevereiro 2014.

CHARD, K. et al. Social Cloud Computing: A Vision for Socially Motivated Resource Sharing. **IEEE TRANSACTIONS ON SERVICES COMPUTING**, v. 5, p. 551-563, 2012.

CHRISTIN, D. et al. Share with strangers: Privacy bubbles as user-centered privacy control for mobile content sharing applications. **Information Security Technical Report**, 2013.

COLUMBUS, L. By 2018, 62% Of CRM Will Be Cloud-Based, And The Cloud Computing Market Will Reach $127.5B. **Forbes**, 2015. Disponivel em: <http://www.forbes.com/sites/louiscolumbus/2015/06/20/by-2018-62-of-crm-will-be-cloud-based-and-the-cloud-computing-market-will-reach-127-5b/>.Acesso em: 10 Agosto 2014

CRN. 6 Revealing Cloud Storage Statistics, 2013. Disponivel em: <http://www.crn.com/slide-shows/cloud/240148574/6-revealing-cloud-storage-statistics.htm>. Acesso em: 10 Julho 2015.

D., M. et al. A study on data deduplication in HPC storage systems. **International Conference for High Performance Computing, Networking, Storage and Analysis (SC)**, 2012. 1-11.

DEEPAK, M.; SHARMA, S. Comprehensive study of data de-duplication. **International Conference on Cloud, Big Data and Trust**, 2013.

DINIZ, T. et al. Integrando o Openstack Keystone com Federações de Identidade. **XIII Simpósio Brasileiro em Segurança da Informação e Sistemas Computacionais, Workshop de Gestão de Identidades Digitais**, 2013. 465-474.

DROPBOX INC. Dropbox, 2015. Disponivel em: <https://www.dropbox.com/>. Acesso em: 10 Julho 2015.

EDUGATE Federation, 2014. Disponivel em: <http://www.edugate.ie/>. Acesso em: 10 Julho 2015.

EMC2. 2011 Digital Universe Study: Extracting Value from Chaos, 2012. Disponivel em: <http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf>. Acesso em: 10 Julho 2015.

ESTEVES, R. A Taxonomic Analysis of Cloud Computing. **1st Doctoral Workshop in Complexity Sciences ISCTE-IUL/FCUL**, 2011.

EUROSTATS. Internet and cloud services - statistics on the use by individuals. **Eurostats - Statistcs Explained**, 2014. Disponivel em: <http://ec.europa.eu/eurostat/statistics-explained/index.php/Internet_and_cloud_services_-_statistics_on_the_use_by_individuals>. Acesso em: Julho 2015.

FACEBOOK , 2012. Disponivel em: <www.facebook.com/>. Acesso em: 10 Julho

2015.

FACEBOOK Graph API Reference User, 2014. Disponivel em: <https://developers.facebook.com/docs/graph-api/reference/v2.0/user>. Acesso em: 10 Julho 2015.

FRAMINGHAM,. Demand from Public Cloud Service Providers and Private Cloud Adopters Will Drive Strong Growth for Full Range of Storage Solutions, According to IDC. **IDC Analyze the Future**, 2011. Disponivel em: <http://www.idc.com/getdoc.jsp?containerId=prUS23097611#.UTaHfjCkqgQ>. Acesso em: 20 Maio 2015.

FRAMINGHAM,. Demand from Public Cloud Service Providers and Private Cloud Adopters Will Drive Strong Growth for Full Range of Storage Solutions, According to IDC. **IDC Analyze the Future**, 2011. Disponivel em: <http://www.idc.com/getdoc.jsp?containerId=prUS23097611#.UTaHfjCkqgQ>. Acesso em: 10 Julho 2015.

FRAMINGHAM, M. Demand from Public Cloud Service Providers and Private Cloud Adopters Will Drive Strong Growth for Full Range of Storage Solutions, According to IDC, 2011. Disponivel em: <http://www.businesswire.com/news/home/20111020005151/en/Demand-Public-Cloud-Service-Providers-Private-Cloud#.UvTotk1dW8Q>. Acesso em: 10 Julho 2015.

FURHT, B.; ESCALANTE, A. J. **Handbook of Cloud Computing**. [S.l.]: [s.n.], 2010.

FUTUREGRID. **Future Grid Portal**, 2012. Disponivel em: <https://portal.futuregrid.org/>. Acesso em: 22 Maio 2015.

GARTNER. Gartner Identifies the Top 10 Strategic Technology Trends for 2013. **Gartner**, 2012. Disponivel em: <http://www.gartner.com/newsroom/id/2209615>. Acesso em: 10 Julho 2015.

GARTNER, 2015. Disponivel em: <http://www.gartner.com/newsroom/id/2867917>. Acesso em: 10 Julho 2015.

GOLLU, K.; SAROIU, S.; WOLMAN, A. A Social Networking-Based Access Control Scheme for Personal Content. **21st ACM Symposium on Operating Systems Principles**, 2007.

GOOGLE. Google. **Google.com**, 2015. Disponivel em: <https://support.google.com/a/answer/2891389?hl=pt-BR>. Acesso em: Junho 2015.

GOOGLE App Engine, 2012. Disponivel em: <https://appengine.google.com/>. Acesso em: 10 Julho 2015.

GOOGLE Docs, 2012. Disponivel em: <docs.google.com/>. Acesso em: 10 Julho 2015.

GOOGLE INC. Google Drive, 2015. Disponivel em: <https://drive.google.com/>.

Acesso em: 10 Julho 2015.

GORDON , R. et al. **High-Tech Tuesday Webinar: Gartner Worldwide IT Spending Forecast, 2Q12 Update: Cloud Is the Silver Lining**. Gartner. [S.l.]. 2012.

GRIFFIN, A. China's great firewall gets higher: tools to evade surveillance and site bans are blocked as Chinese internet censors tighten grip. **The Independent**, 2015. Disponivel em: <http://www.independent.co.uk/life-style/gadgets-and-tech/news/chinas-great-firewall-gets-higher-tools-to-evade-surveillance-and-site-bans-are-blocked-as-chinese-internet-censors-tighten-grip-10013537.html>. Acesso em: 07 Junho 2015.

HARDT, D. **The OAuth Autorization Framework**. [S.l.]. 2012.

HARDT, D. The OAuth 2.0 Authorization Framework, 2013. Disponivel em: <https://www.rfc-editor.org/rfc/rfc6749.txt>. Acesso em: 10 Fevereiro 2015.

INCOMMON Federation, 2014. Disponivel em: <http://www.incommonfederation.org/>. Acesso em: 7 Julho 2014.

JIN, K.; MILLER, E. The Effectiveness of Deduplication on Virtual Machine Disk Images. **SYSTOR '09 Proceedings of SYSTOR 2009: The Israeli Experimental Systems Conference**, 2009.

JOHN, K.; BUBENDORFER, K.; CHARD, K. A Social Cloud for Public eResearch. **International Conference on eScience**, 2011.

JøSANG, A.; POPE, S. User Centric Identity Management. **AusCERT Conference**, 2005.

KAANICHE, N.; LAURENT, M. A Secure Client Side Deduplication Scheme in Cloud Storage Environments. **New Technologies, Mobility and Security (NTMS), 2014 6th International Conference on**, Dubai, 4 Abril 2014. 1-7.

KHAN, A. Access Control in cloud Enviroment. **ARPN Journal of Engineering and Applied Sciences** , 2012. ISSN 1819-6608.

KOSHY, J.; BUBENDORFER, K.; CHARD, K. A Social Cloud for Public E-research. **E-Science (e-Science), 2011 IEEE 7th International Conference on**, p. 363-370, 2011.

KOULOUZIS, S. et al. Cloud Data Federation for Scientific Applications. **Euro-Par 2013: Parallel Processing Workshops**, v. 8374, p. 13-22, 2014.

LOWE, S. Data deduplication is an increasingly important aspect of storage technology. **Wikibon**, 2012. Disponivel em: <http://wikibon.org/blog/data-deduplication-is-an-increasingly-important-aspect-of-storage-technology/>. Acesso em: 10 Fevereiro 2015.

MAJUMDER, ; NAMASUDRA, ; NATH, S. Taxonomy and Classification of Access

Control Models for Cloud Environments. **Continued Rise of the Cloud**, p. 23-53, 2014.

MCKENDRICK, J. Cloud Computing's Hidden 'Green' Benefits. **Forbes**, 2011. Disponivel em: <http://www.forbes.com/sites/joemckendrick/2011/10/03/cloud-computings-hidden-green-benefits/>. Acesso em: 10 Fevereiro 2015.

MEISTER, D.; BRINKMANN, A. Multi-level comparison of data deduplication in a backup scenario. **he Israeli Experimental Systems Conference**, 2009.

MELL, P.; GRANCE, T. **The NIST Definition of Cloud Computing**. NIST. [S.l.]. 2011.

MEMOPAL. Memopal Inc., 2015. Disponivel em: <http://www.memopal.com/pt-br/>. Acesso em: 10 Julho 2015.

MIRCORSOFT. Skidrive, 2015. Disponivel em: <https://skydrive.live.com/>. Acesso em: 10 Julho 2015.

MOHAMED, A. A history of cloud computing. **ComputerWeekly.com**, 2012. Disponivel em: <http://www.computerweekly.com/feature/A-history-of-cloud-computing>. Acesso em: 10 Agosto 2014.

MOZY INC. Mozy, 2015. Disponivel em: <http://mozy.com/>. Acesso em: 10 Julho 2015.

PUNCEVA, M. et al. Incentivising Resource Sharing in Social Clouds. **Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), IEEE 21st International Workshop on**, 2012. 185-190.

QUANTUM IQ. BIG DATA: Managing Explosive Growth, 2012. Disponivel em: <https://www.google.com.br/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&ved=0CEUQFjAA&url=https%3A%2F%2Fiq.quantum.com%2FexLink.asp%3F8737629OB17F91I30026979&ei=SuP0Uo-2CMf0kQelhICwCA&usg=AFQjCNG4lrLWow1M776oJkpF8TU_-92TEw&sig2=CY8ICSmCUnokayQmNqpa_g&b>. Acesso em: 07 Fevereiro 2014.

RACKSPACE , 2015. Disponivel em: <http://www.rackspace.com/>. Acesso em: 10 Agosto 2015.

RNP. RNP no mapa das federações de identidade mundiais. **Portal RNP**, 2015. Disponivel em: <http://portal.rnp.br/web/rnp/noticias/-/rutelistaconteudo/RNP-no-mapa-das-federacoes-de-identidade-mundiais/495061_o80B;jsessionid=F49FBBE17FF093DF50680B351F92BEFD.inst1>. Acesso em: Maio 2015.

SAML Especifications, 2014. Disponivel em: <http://saml.xml.org/saml-specifications>. Acesso em: 9 Julho 2014.

SHIBBOLETH , 2014. Disponivel em: <https://shibboleth.net/>. Acesso em: 8 Julho 2015.

SILVA, E.; MENEZES, E. **Metodologia da Pesquisa e Elaboração de Dissertação**. 4ª edição revisada e atualizada. ed. [S.l.]: Universidade Federal de Santa Catarina - UFSC, 2001.

SILVA, et al. Caixas de Interesses: um Novo Mecanismo para a Colaboração através de Nuvem de Armazenamento de Dados. **Anais SBSC 2014**, Curitiba, Outubro 2014.

SILVA, L. et al. Estudo de caso: Integração de Clientes de Nuvem Openstack Swift com Federação de Identidade. **XIII Simpósio Brasileiro em Segurança da Informação e Sistemas Computacionais, Workshop de Gestão de Identidades Digitais**, 2013. 455-464.

SIMPLESAMLPHP , 2014. Disponivel em: <https://simplesamlphp.org/>. Acesso em: 8 Julho 2014.

THAUFEEG, A.; BUBENDORFER, K.; CHARD, C. A Social Cloud for Public eResearch. **International Conference on eScience**, 2011.

THAUFEEG, A.; BUBENDORFER, K.; CHARD, K. A Collaborative E-research in a Social Cloud. **E-Science (e-Science), 2011 IEEE 7th International Conference on**, p. 224-231, 2011.

VISWANATHAN, P. Cloud Computing – Is it Really All That Beneficial?. **About.com**, 2012. Disponivel em: <http://mobiledevices.about.com/od/additionalresources/a/Cloud-Computing-Is-It-Really-All-That-Beneficial.htm>. Acesso em: 20 evereiro 2013.

WAINER, J. **Métodos de pesquisa quantitativa e qualitativa para a Ciência da Computação**. Rio de Janeiro: [s.n.], 2007.

WANGHAM, M. et al. Gerenciamento de Identidades Federadas. **O Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais - SBSeg**, 2010.

WINDOWS Azure Plataform, 2012. Disponivel em: <http://www.windowsazure.com>. Acesso em: 10 Julho 2015.

YOUNIS, Y. A.; KIFAYAT, K.; MERABTI, M. An access control model for cloud computing. **Journal of Information Security and Applications**, v. 19, p. 45 - 60, 2014.