Caracterização dos Arquivos Armazenados no Dropbox

Idilio Drago[1], Alex Borges Vieira[2], Ana Paula Couto da Silva[3]

[1] University of Twente
[2]DCC - Universidade Federal de Juiz de Fora
[3]DCC - Universidade Federal de Minas Gerais

i.drago@utwente.nl; alex.borges@ufjf.edu.br; ana.coutosilva@dcc.ufmg.br

Resumo. *Aplicações de armazenamento em nuvem (cloud storage) tem se tornado cada vez mais populares. Tal sucesso se deve à facilidade, praticidade e segurança oferecidas por estes sistemas, o que atrai tanto usuários domésticos quanto empresas. Apesar desta crescente popularidade, ainda são raras as pesquisas caracterizando o uso e a carga típica destes serviços. Esse trabalho apresenta uma caracterização preliminar dos arquivos armazenados no Dropbox, atualmente o sistema de cloud storage mais utilizado no mundo. Tal caracterização se baseia em uma coleta de dados de mais de 300 voluntários em diversos países. Os resultados encontrados mostram que a maioria dos arquivos são binários, que normalmente não são modificáveis. Isso sugere que os mecanismos de controle de versão e atualização diferencial implementados pelo Dropbox causam uma redução desprezível no tráfego da aplicação. Por outro lado, há também um número significante de réplicas, que são transferidas pelo Dropbox apenas uma vez. Espera-se que o presente trabalho seja um ponto de partida para a geração de cargas sintéticas realistas e, consequentemente, para o desenvolvimento de sistemas de cloud storage mais robustos.*

Palavras-chave: Cloud storage, Dropbox.

1. Introdução

Computação nas nuvens [Zhang et al. 2010], ou *cloud computing*, tem atraído cada vez mais a atenção da indústria e da academia. Atualmente, há uma crescente oferta de serviços baseado nesse paradigma, o que aumenta o volume de dados gerados por tais aplicações. Neste contexto, serviços de armazenamento em nuvem (*cloud storage*) vem ganhando destaque. Tanto empresas quanto usuários domésticos podem armazenar seus dados remotamente de maneira simples, prática e segura. Tal popularização tornar-se mais evidente pelo ingresso de grandes provedores, como Google e Microsoft, neste setor.

Apesar da popularidade, há uma lacuna de trabalhos caracterizando o uso e a carga típica destes serviços. Informações sobre o funcionamento de tais aplicações são raras, uma vez que os conteúdos armazenados são privados e as soluções de sincronizações são, majoritariamente, proprietárias. Além disso, a coleta de dados sobre o uso destes sistemas é particularmente desafiadora, dado que as aplicações usam protocolos criptografados.

Assim, atualmente ainda não é possível caracterizar precisamente as aplicações de *cloud storage* disponíveis. Conhecer as características tanto dos sistemas, quanto de seus usuários, é indispensável para o planejamento dos recursos necessários ao provimento destes serviços com qualidade. Isso possibilitará o desenvolvimento de sistemas de *cloud storage* robustos, tolerantes a falhas e de melhor desempenho.

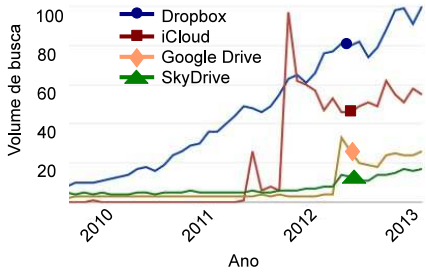

Figura 1. Interesse de busca por aplicações de *cloud storage*.

Este artigo apresenta uma caracterização preliminar dos arquivos armazenados no Dropbox (`http://www.dropbox.com`), o sistema de *cloud storage* mais utilizado atualmente no mundo. Tal caracterização é realizada a partir de dados enviados por 353 voluntários de vários países. Os resultados mostram que a grande parte dos arquivos salvos pelos usuários do Dropbox são binários normalmente não-editáveis, como imagens JPEGs, documentos PDFs e programas executáveis. Isso sugere que os mecanismos de controle de versão e atualização diferencial do Dropbox não reduzem significativamente o tráfego da aplicação. Porém, um número considerável de réplicas também é observado, o que permite ao cliente Dropbox enviar e armazenar as cópias uma única vez.

O presente trabalho é pioneiro em caracterizar os arquivos dos usuários do Dropbox. O protocolo proprietário desta aplicação, bem como várias características de seu tráfego, são analisados em [Drago et al. 2012]. Esta pesquisa, entretanto, não fornece informações sobre os arquivos armazenados no serviço. Outros trabalhos comparam provedores de *cloud storage* ou focam no desempenho dos sistemas [Hu et al. 2010, Bergen et al. 2011], sem considerar contudo o possível efeito causado por diferentes tipos de arquivos. Finalmente, vários autores investigam aspectos relacionados a segurança e privacidade em *cloud storage* [Ion et al. 2011, Mulazzani et al. 2011, Halevi et al. 2011], independentemente do uso ou dos tipos de arquivos salvos nas aplicações.

O restante deste artigo está assim organizado: A Seção 2 apresenta um resumo do funcionamento do Dropbox; a Seção 3 descreve a metodologia utilizada na coleta de dados. Em seguida, os resultados preliminares são discutidos na Seção 4. Por fim, a Seção 5 sumariza as contribuições do trabalho e indica futuras direções.

2. A Aplicação Dropbox

O Google Trends (`http://www.google.com/trends/`) sugere que o Dropbox é o serviço de *cloud storage* mais popular no mercado. A Figura 1 nos mostra que desde 2010 o Dropbox atraiu a maior parte das buscas no Google, se comparado a seus maiores concorrentes. De acordo com as análises conduzidas em [Drago et al. 2012], o Dropbox é responsável por cerca de 4% do volume tráfego em algumas redes. Este volume corresponde a cerca de 30% do tráfego gerado pelo YouTube, um dos sistemas mais populares de distribuição de vídeo na Internet.

O serviço fornecido pelo Dropbox baseia-se no armazenamento dos arquivos de seus usuários em servidores com alta disponibilidade. Há dois componentes principais na arquitetura do Dropbox [Drago et al. 2012]. O primeiro refere-se aos servidores de controle da aplicação, e são mantidos diretamente pela empresa. O segundo componente refere-se aos servidores de armazenamento de dados, e são hospedados pela Amazon (`http://aws.amazon.com`). Em ambos os casos, subdomínios de *dropbox.com*

permitem diferenciar as partes do serviço que executam funcionalidades específicas.

Dentre outras funcionalidades, os usuários do Dropbox podem sincronizar vários dispositivos através de uma mesma conta. Os usuários também são capazes de sincronizar arquivos seletivamente, assim como controlar os recursos de rede utilizados pelo cliente. O acesso ao serviço é fornecido através de um aplicativo com versões nativas para Windows, Linux e Mac, além de acesso através de uma interface *web*. O Dropbox também provê interfaces de programação para que desenvolvedores criem aplicações para diversos ambientes, como sistemas móveis Android.

Durante a transferência de dados entre clientes e servidores Dropbox, observa-se uma considerável redução no tamanho dos arquivos [Hu et al. 2010]. De fato, todos os dados são compactados ainda na estação cliente, com o objetivo de reduzir, por consequência, o tempo total da transferência. Além disso, o cliente Dropbox compara versões de um mesmo arquivo, transferindo apenas as suas diferenças. Ainda, arquivos duplicados de um mesmo usuário são transferidos apenas uma vez. A eficácia desses três mecanismos, porém, varia de acordo com os tipos de arquivo salvo no sistema. Por fim, todos os dados trocados são criptografados, satisfazendo condições básicas de privacidade e segurança. O protocolo *HTTPS* é utilizado para acessar a maior parte dos servidores.

3. Metodologia de Coleta dos Dados

A caracterização apresentada dos arquivos armazenados no Dropbox é baseada em coletas realizadas a partir de voluntários. A chamada para participação foi direcionada a toda comunidade, mas houve uma maior adesão de voluntários no Brasil e na Europa. Mais ainda, a grande maioria dos voluntários utilizam Dropbox com fins acadêmicos. Em outras palavras, são alunos ou pesquisadores de universidades.

Durante a coleta de informações, os voluntários executaram um programa desenvolvido pelo grupo com a finalidade de buscar as principais características dos arquivos armazenados no Dropbox. Os voluntários também respondiam a um formulário com perguntas a respeito de seu perfil. Cerca de 88% dos voluntários são homens com idade entre 20 e 30 anos. Apenas 4,5% dos voluntários declararam pagar pelo uso de *cloud storage*. A capacidade de armazenamento média declarada é de 23,4 GB.

O programa de coleta de informações foi desenvolvido em versões nativas para Windows, Mac e Linux, além de uma versão especial para a plataforma Java. O programa de coleta, inicialmente, lê as informações básicas sobre o sistema Dropbox instalado, como o diretório padrão de armazenamento da aplicação. A seguir, o programa varre recursivamente a pasta inicial do Dropbox. Todas as informações coletadas são anonimizadas de tal forma que o conteúdo e o voluntário não possam ser identificados.

Mais precisamente, são coletados metadados sobre o processo de coleta de informações, como o tempo inicial e o tempo final de captura. Para cada voluntário, é associado um identificador numérico único, permitindo que um mesmo voluntário contribua mais de uma vez com a coleta. Para cada arquivo encontrado, são armazenadas as informações sobre o tamanho, a extensão e o tipo *MIME* do arquivo. Também é capturada a data de última alteração do arquivo ou pasta analisada.

Cada arquivo analisado na coleta é identificado através de uma chave composta pela chave *Hash* dos 8 kB iniciais e dos 8 kB finais do arquivo. Nesse sentido, assume-se que se dois ou mais arquivos diferentes tiverem a mesma chave composta, mesmo tamanho e mesmo tipo MIME, eles são réplicas. Tal abordagem simplificada de

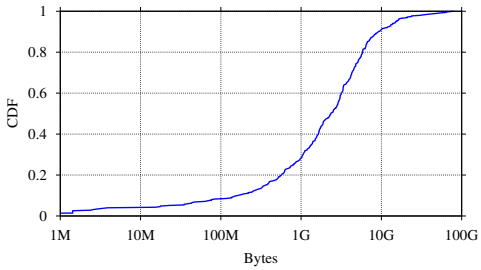

Figura 2. Distribuição do espaço ocupado por usuário.

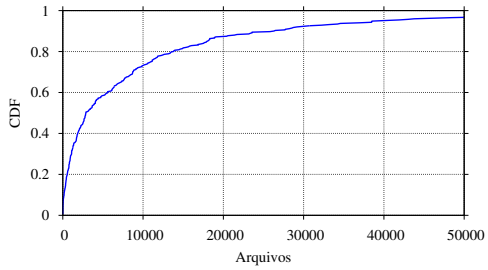

Figura 3. Distribuição do número de arquivos por usuário.

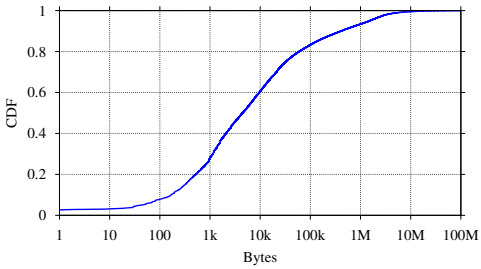

Figura 4. Distribuição do tamanho dos arquivos.

identificação de arquivo foi utilizada principalmente para reduzir o tempo de coleta (por não varrer o arquivo por completo para o cálculo da chave *Hash*).

Ao fim da coleta, os dados eram apresentados aos voluntários e enviados a um servidor centralizado. Tais dados são encaminhados para a análise somente após a aprovação explícita do participante do experimento. Nesse momento, os voluntários podiam também acessar suas estatísticas básicas e verificar, entre outras informações, a distribuição de tipo de arquivos que ele armazena em seu repositório Dropbox.

Nas análises apresentadas, são consideradas apenas dados referentes a usuários únicos. Caso um voluntário tenha enviado suas estatísticas mais de uma vez, é considerada a última postagem. O conjunto de dados avaliado nesse trabalho contém 420 coletas de 353 usuários únicos. Foram avaliados mais de 1,4TB de arquivos do Dropbox. Cerca de 45% de voluntários são da América Latina, 7% América do Norte e 40% da Europa.

4. Caracterização Preliminar dos Arquivos Armazenados no Dropbox

Nessa seção são apresentadas as características básicas dos arquivos armazenados no Dropbox dos voluntários. Os resultados apresentados discutem o perfil dos usuários do Dropbox e servem para direcionar futuros trabalhos na área. Mais ainda, os resultados apresentados servem como um ponto de partida para o planejamento de capacidade de serviços de armazenamento com características semelhantes ao Dropbox.

A maior parte dos voluntários apresenta uma grande quantidade de dados armazenados em seu Dropbox. A Figura 2 apresenta a distribuição de probabilidade acumulada do consumo de espaço em disco pelos usuários do Dropbox. Observa-se que em mais de 70% dos casos, os usuários armazenam pelo menos 1GB de dados. Mais ainda, praticamente 10% deles armazena mais que 10GB em seus repositórios.

De acordo com a Figura 3, a maior parte dos usuários contém menos de 10 mil arquivos no Dropbox. Porém, há uma quantia significativa de usuários com um grande número de arquivos em disco. De fato, cerca de 10% dos usuários tem quantias acima de 20 mil arquivos. Os arquivos armazenados no Dropbox apresentam um tamanho

médio na ordem de 1 kB. A Figura 4 apresenta a distribuição dos tamanhos dos arquivos encontrados no Dropbox. Percebe-se que 90% dos arquivos são menores que 1MB. A quantidade absoluta de arquivos maiores que 10MB é praticamente desprezível.

A Figura 5 mostra o volume de dados que cada tipo de arquivo ocupa no Dropbox (em % do total de bytes de todos os arquivos de todos os voluntários). A maior parte dos arquivos manuseados por usuários do Dropbox são arquivos não comumente editáveis, como imagens JPEG, documentos PDF, vídeo MPEG, e outros arquivos binários. Por exemplo, praticamente 20% do volume de dados correspondem a imagens. Nesse caso, o uso de controle de versão e atualizações por diferença de conteúdo do Dropbox pode não apresentar um bom desempenho. Isso porque, a cada alteração do arquivo, é normal que a maior parte da imagem (ou dos outros formatos binários) seja alterada.

A proporção do número de arquivos é similar a distribuição encontrada para o espaço ocupado por cada tipo de arquivo. De acordo com a Figura 6, cerca de 7% dos arquivos são imagens do tipo JPEG. Apesar dos arquivos de vídeo impactarem o espaço em disco, eles não figuram entre os 15 tipos de arquivos mais populares.

Finalmente, a Figura 7 e a Figura 8 mostram que há uma quantidade não desprezível de arquivos replicados no Dropbox dos voluntários. Por exemplo, em mais de 20% dos casos, os usuários apresentam mais de 17% do seu espaço no Dropbox consumido por arquivos replicados. Em números absolutos, em mais de 20% dos casos, pelo menos 40% dos arquivos (independente de seu tamanho) são replicados. Durante a criação ou atualização, réplicas são transferidas pelo Dropbox apenas uma vez e, assim, não consomem banda de rede de forma desnecessária.

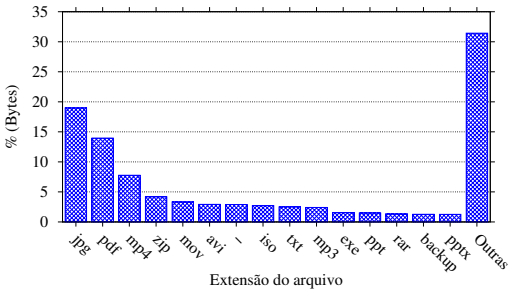

Figura 5. Espaço ocupado por cada tipo de arquivos.

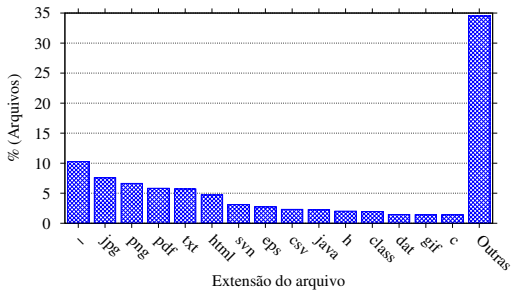

Figura 6. Quantidade de arquivos de acordo com a extensão.

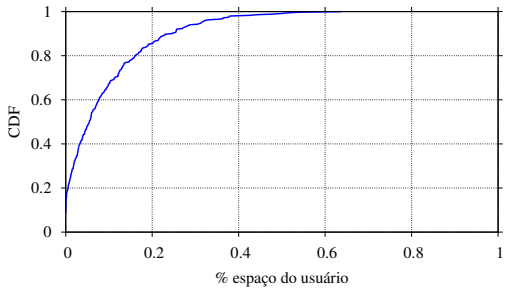

Figura 7. Consumo de espaço pelas réplicas Dropbox.

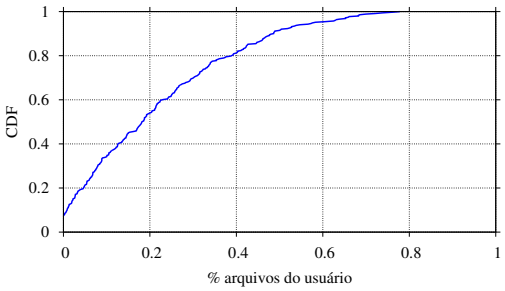

Figura 8. Quantidade de arquivos replicados.

5. Conclusões e Trabalhos Futuros

Cloud computing tem atraído uma crescente atenção da indústria e da academia. Apesar da popularidade dessas aplicações, ainda são raras as caracterizações sobre elas.

Portanto, esse artigo apresenta uma caracterização preliminar dos arquivos armazenados no Dropbox, a aplicação mais popular de *cloud storage* no mercado.

Os resultados encontrados mostram que a maioria dos arquivos encontrados no Dropbox não são normalmente modificáveis. Por exemplo, mais de 30% dos arquivos são imagens, músicas ou vídeos. Esses tipos de arquivos podem fazer com que os mecanismos de controle de versão e atualização diferencial implementados pelo Dropbox tenham um impacto desprezível na economia tráfego de gerado pela aplicação. Por outro lado, há também um número significante de réplicas. Réplicas são transferidas pelo Dropbox apenas uma vez e, assim, não incorrem em um maior consumo de banda de rede.

Finalmente, espera-se que o presente trabalho seja um ponto de partida para a geração de cargas sintéticas realistas e, consequentemente, para o desenvolvimento de sistemas de *cloud storage* mais robustos. Como principal trabalho futuro, espera-se expandir a presente caracterização com informações sobre como os usuários criam e atualizam seus arquivos ao longo de uma sessão de uso do Dropbox. Mais ainda, serão comparados os perfis de arquivos armazenados em *cloud storage* com outros sistemas de arquivo conhecidos [Douceur and Bolosky 1999, Agrawal et al. 2007].

6. Agradecimentos

This work was partially funded by CNPq, CAPES, Fapemig. Thanks to all Dropbox volunteers. Thanks to Marco Melia. Thanks to the EU-IP project mPlane. The mPlane is funded by the European Commission under the grant n-318627.

Referências

Agrawal, N., Bolosky, W. J., Douceur, J. R., and Lorch, J. R. (2007). A Five-Year Study of File-System Metadata. *ACM Transactions on Storage*, 3(3).

Bergen, A., Coady, Y., and McGeer, R. (2011). Client Bandwidth: The Forgotten Metric of Online Storage Providers. In *Proceedings of the 2011 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, PacRim'2011.

Douceur, J. R. and Bolosky, W. J. (1999). A Large-Scale Study of File-System Contents. *SIGMETRICS Perform. Eval. Rev.*, 27(1):59–70.

Drago, I., Mellia, M., Munafò, M. M., Sperotto, A., Sadre, R., and Pras, A. (2012). Inside Dropbox: Understanding Personal Cloud Storage Services. In *Proceedings of the 12th ACM Internet Measurement Conference*, IMC'12, pages 481–494.

Halevi, S., Harnik, D., Pinkas, B., and Shulman-Peleg, A. (2011). Proofs of Ownership in Remote Storage Systems. In *Proceedings of the 18th ACM Conference on Computer and Communications Security*, CCS'11, pages 491–500.

Hu, W., Yang, T., and Matthews, J. N. (2010). The Good, the Bad and the Ugly of Consumer Cloud Storage. *ACM SIGOPS Operating Systems Review*, 44(3):110–115.

Ion, I., Sachdeva, N., Kumaraguru, P., and Čapkun, S. (2011). Home is safer than the cloud!: privacy concerns for consumer cloud storage. In *Proceedings of the Seventh Symposium on Usable Privacy and Security*, page 13.

Mulazzani, M., Schrittwieser, S., Leithner, M., Huber, M., and Weippl, E. (2011). Dark Clouds on the Horizon: Using Cloud Storage as Attack Vector and Online Slack Space. In *Proceedings of the 20th USENIX Conference on Security*, SEC'11.

Zhang, Q., Cheng, L., and Boutaba, R. (2010). Cloud Computing: State-of-the-Art and Research Challenges. *Journal of Internet Services and Applications*, 1:7–18.