Avaliação do Controle de Acesso de Múltiplos Usuários a Múltiplos Arquivos em um Ambiente Hadoop

Eduardo Scuzziato[1], João E. Marynowski[1,2], Altair O. Santin[1]

[1]Escola Politécnica – Ciência da Computação
Pontifícia Universidade Católica do Paraná (PUCPR), Curitiba – PR – Brasil

[2]Setor de Educação Profissional e Tecnológica
Universidade Federal do Paraná (UFPR), Curitiba – PR – Brasil

scuzziato.eduardo@gmail.com, jeugenio@ufpr.br, santin@ppgia.pucpr.br

Abstract. *Massive processing of data is a reality for several computer systems. The security of processed data has great importance since the environment is typically shared among multiple users. This article presents an evaluation of the access control of multiple users and multiple files, considering the different control levels of a Hadoop environment (operating system, distributed file system and web interface). A test scenario is proposed and validated at different levels and different versions of a Hadoop distribution (Hortonworks). The versions presented the same behavior but we identified errors and differences between control levels.*

Resumo. *O processamento massivo de dados é uma realidade para diversos sistemas computacionais. A segurança dos dados processados é de grande importância, uma vez que o ambiente normalmente é compartilhado entre múltiplos usuários. Este artigo apresenta uma avaliação do controle de acesso de múltiplos usuários a múltiplos arquivos, considerando os diferentes níveis de controle de um ambiente Hadoop (sistema operacional, sistema de arquivo distribuído e interface web). Um cenário de teste é proposto e validado nos diferentes níveis e diferentes versões de uma distribuição do Hadoop (Hortonworks). As versões apresentaram mesmo comportamento mas identificamos erros e diferenças entre os níveis de controle.*

1. Introdução

Conjuntos de dados extremamente grandes são gerados e precisam ser manipulados diariamente excedendo a capacidade de processamento dos sistemas de banco de dados convencionais [Zikopoulos et. al. 2010]. Esses conjunto de dados são denominados big data e têm como características o grande volume, velocidade e a variabilidade dos dados [White 2012]. Hadoop é um framework de código aberto que permite o armazenamento e processamento distribuído de big data em um grande conjunto de máquinas (*cluster*) [Hadoop]. Nesse ambiente, o armazenamento é feito pelo HDFS (Hadoop Distributed File System) que é um sistema de arquivos distribuído escalável para grandes aplicações e com grande quantidade de dados distribuídos [HDFS]. O Hadoop também dispõe do Hue (Hadoop User Experience) [Hue], que é uma aplicação web que fornece aos seus usuários uma interface para a manipulação de outras ferramentas do ambiente Hadoop,

como o HDFS, Hive [Thusoo 2009] e Hbase [White 2012]; além da administração de usuários, que possibilita criar usuários e grupos.

Em um ambiente que possibilita diversas formas de uso e dispõe de grande poder de processamento como Hadoop, é natural a ideia de que ele possa ser compartilhado entre diferentes usuários e grupos [Tankard 2012]. Essa possibilidade de compartilhamento faz com que questões de segurança, como a restrição de acesso a determinados diretórios ou arquivos, sejam essenciais. As possibilidades de compartilhamento podem ser diferenciadas segundo os diferentes níveis de controle: sistema operacional (SO), sistema de arquivo distribuído (HDFS) e pela interface web (Hue).

Considerando os níveis de acesso e controle de arquivos, é importante realizar experimentos para verificar a coerência e o correto funcionamento em todos os níveis. Usualmente, testes unitários são empregados durante e após o desenvolvimento de sistemas Hadoop [Hadoop, HDFS, Hive, White 2012, Tabatabaei 2014]. Entretanto, diversos problemas referentes ao mal funcionamento e ineficiências relacionadas a segurança dos dados acabam sendo relatados por usuários durante a utilização [HadoopIssues, HDFSIssues, Tankard 2012, Bertino 2015].

Este artigo apresenta uma avaliação do controle de acesso de múltiplos usuários a múltiplos arquivos em um ambiente Hadoop envolvendo três níveis de controle e três versões da uma distribuição Hadoop. Um método de avaliação e validação é apresentado, a fim de verificar a segurança das informações considerando regras de restrição para múltiplos usuários, grupos e diversos arquivos com diferentes restrições de acesso. Experimentos são apresentados e identificamos comportamentos divergentes e errados das versões, e principalmente quando usado a interface web Hue, pela qual o comportamento foi o mesmo em todas as versões, porém, com erros.

2. Método

O método para a avaliação do controle de acesso e manipulação de arquivos foi realizado a partir de um cenário que contemplasse as diferentes formas de acesso aos arquivos. Foi criado um cenário no qual um conjunto de usuários foi organizado em grupos de modo que se pudesse avaliar o comportamento no controle de acesso de múltiplos arquivos. O objetivo dos experimentos é validar se as restrições de acesso são válidas em três níveis (SO, HDFS e Hue), considerando permissões de acesso para usuários, grupos e outros. O cenário contempla apenas permissão de leitura e escrita, já que a execução não se aplica a arquivos de dados.

A Figura 1 representa a organização de usuários e grupos no cenário proposto. São três usuários (USUARIO1, USUARIO2 e USUARIO3) e dois grupos (GRUPO1 e GRUPO2). O USUARIO1 e o USUARIO2 pertencem ao GRUPO1, e o USUARIO3 ao GRUPO2. Dessa forma, levamos em consideração as restrições que podem ser implementadas para arquivos e diretórios em relação a usuários, grupos e outros. Dessa forma é possível validar a restrição de acesso de múltiplos usuários a múltiplos arquivos.

A Figura 2 representa a forma como estão organizados os arquivos dentro do cenário, considerando a localização e tipos de restrições impostas a diretórios e arquivos. O diretório USUARIO1 permite o acesso total somente ao USUARIO1, já o

diretório GRUPO1 permite o acesso total somente aos usuários pertencentes ao GRUPO1. Um conjunto de arquivos de texto foi criado dentro de cada diretório USUARIO1 e GRUPO1, com restrições de somente leitura, somente escrita; e leitura e escrita apenas para o USUARIO1 e integrantes do GRUPO1.

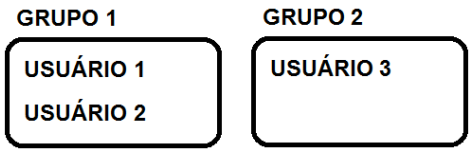

Figura 1. Distribuição de usuários por grupo.

Figura 2. Distribuição de arquivos e diretórios.

3. Experimentos e Avaliação

Os experimentos foram realizados usando três versões da distribuição Hadoop da Hortonworks - HDP (HortonWorks Data Plataform): 1.3, 2.1 e 2.2.4. O principal motivo da comparação entre as três versões é a validação do comportamento dos respectivos componentes em cada versão, as quais são:

- HDP 1.3: Hadoop 1.2.0, HDFS 1.2.0, Hue 2.2.0 e CentOS 2.6.32-358.
- HDP 2.1: Hadoop 2.4.0, HDFS 2.4.0, Hue 2.5.1 e CentOS 2.6.32-431.
- HDP 2.2.4: Hadoop 2.6.0, HDFS 2.6.0, Hue 2.6.1 e CentOS 2.6.32-504.

3.1. Execução do Método

Em cada versão do HDP foi implementado o cenário nos respectivos SO, HDFS e Hue. Ou seja, o cenário foi implementado e executado 9 vezes, objetivando eliminar possíveis desvios de comportamento, mas que não ocorreram. Os grupos, usuários e arquivos foram criados nos diferentes diretórios seguindo as restrições de acesso conforme definido pelo método (Seção 2).

Os usuários, diretórios e arquivos foram manipulados no Linux utilizando os comandos: *adduser*, *groupadd*, *usermod*, *mkdir*, *chmod*, *chgrp*, *cd* e *cat*. O HDFS utiliza os usuários e grupos do próprio SO, logo, os comandos utilizados foram os mesmos para administrar usuário no Linux. Mas, para manipular os arquivos foram utilizados os comandos específicos do HDFS, que basicamente apenas adicionam o "-" antes do comando Linux, como *–mkdir* e *–chmod*. Já no Hue, foram utilizadas as opções disponibilizadas nos menus de gerenciamento de usuários (*Hue Users*) e gerenciamento de arquivos (*File Browser)*.

3.2. Resultados

Em todas as versões do HDP (1.3, 2.1 e 2.2.4) tanto o SO quanto o HDFS asseguraram as restrições impostas aos usuários e grupos. Foram garantidas as restrições de acesso aos diretórios e arquivos, assim como as restrições de operações (leitura e escrita). Esses resultados são justificados uma vez que os usuários, grupos e restrições de acesso do HDFS são manipulados utilizando o SO.

Por outro lado, o Hue foi reprovado em diversos testes realizados (Tabela 1). Os usuários tiveram acesso a todos os diretórios e a todos os arquivos, mesmo aqueles com restrições de acesso e operações sobre os arquivos. Esses resultados se devem ao fato de que os usuários, grupos e permissões criadas pelo Hue deveriam ser implementados por uma lista de controle de acesso (ACL) no HDFS, o que não ocorreu. As primeiras versões do HDFS não implementaram ACL e mesmo a última versão testada (2.6.0) também não restringiu o acesso, mesmo criando ACLs.

Tabela 1. Resultados para o Hue.

Usuário	Diretório	Arquivo	Operação	Resposta	Resultado
USUARIO1	/USUARIO1	LEITURA.TXT	LEITURA	Permite	Aprovado
USUARIO1	/USUARIO1	ESCRITA.TXT	LEITURA	Permite	**Reprovado**
USUARIO1	/USUARIO1	LEITURA_ESCRITA.TXT	LEITURA	Permite	Aprovado
USUARIO1	/USUARIO1	LEITURA.TXT	ESCRITA	Permite	**Reprovado**
USUARIO1	/USUARIO1	ESCRITA.TXT	ESCRITA	Permite	Aprovado
USUARIO1	/USUARIO1	LEITURA_ESCRITA.TXT	ESCRITA	Permite	Aprovado
USUARIO2	/ USUARIO1	LEITURA.TXT	LEITURA	Permite	**Reprovado**
USUARIO2	/ USUARIO1	ESCRITA.TXT	LEITURA	Permite	**Reprovado**
USUARIO2	/ USUARIO1	LEITURA_ESCRITA.TXT	LEITURA	Permite	**Reprovado**
USUARIO2	/ USUARIO1	LEITURA.TXT	ESCRITA	Permite	**Reprovado**
USUARIO2	/ USUARIO1	ESCRITA.TXT	ESCRITA	Permite	**Reprovado**
USUARIO2	/ USUARIO1	LEITURA_ESCRITA.TXT	ESCRITA	Permite	**Reprovado**
USUARIO2	/GRUPO1	LEITURA.TXT	LEITURA	Permite	Aprovado
USUARIO2	/GRUPO1	ESCRITA.TXT	LEITURA	Permite	**Reprovado**
USUARIO2	/GRUPO1	LEITURA_ESCRITA.TXT	LEITURA	Permite	Aprovado
USUARIO2	/GRUPO1	LEITURA.TXT	ESCRITA	Permite	**Reprovado**
USUARIO2	/GRUPO1	ESCRITA.TXT	ESCRITA	Permite	Aprovado
USUARIO2	/GRUPO1	LEITURA_ESCRITA.TXT	ESCRITA	Permite	Aprovado
USUARIO3	/GRUPO1	LEITURA.TXT	LEITURA	Permite	**Reprovado**
USUARIO3	/GRUPO1	ESCRITA.TXT	LEITURA	Permite	**Reprovado**
USUARIO3	/GRUPO1	LEITURA_ESCRITA.TXT	LEITURA	Permite	**Reprovado**
USUARIO3	/GRUPO1	LEITURA.TXT	ESCRITA	Permite	**Reprovado**
USUARIO3	/GRUPO1	ESCRITA.TXT	ESCRITA	Permite	**Reprovado**
USUARIO3	/GRUPO1	LEITURA_ESCRITA.TXT	ESCRITA	Permite	**Reprovado**

As falhas ocorreram em todas as versões do HDP, uma vez que a interface usa um usuário administrador no HDFS e todos os usuários manipulam o sistema de arquivos por esse usuário. Assim, o controle que o HDFS possuía, sem ACL, é prejudicado, pois o usuário é único, independentemente do usuário que está registrado e manipulando os diretórios e arquivos no Hue. O controle deveria ocorrer com ACL mas não ocorreu, mesmo ativando a referida funcionalidade através da propriedade "dfs.namenode.acls.enabled" no arquivo de configuração do HDFS (hdfs-site.xml) e reiniciando o sistema.

4. Considerações Finais

Este artigo contribui com um método e um conjunto de experimentos necessários para avaliação do controle de acesso de múltiplos usuários a múltiplos arquivos em um ambiente Hadoop. Também contribui identificando comportamentos divergentes e errados das versões e principalmente quando usado a interface web Hue. Foram testadas diferentes versões da distribuição do Hadoop Hortonworks (HDP), considerando tanto o sistema operacional Linux, quanto o HDFS e a interface web Hue. Um cenário de teste foi criado e implementado considerando esses diferentes níveis de um ambiente Hadoop. Foi verificado que no Linux e HDFS o controle de acesso é realizado como esperado em todas as versões HDP. No entanto, ocorreram diferenças nas versões considerando o Hue. Nas versões HDP sem ACL, o usuário é único e isso inviabiliza o controle de acesso de diferentes usuários. Ativando ACL, as restrições de acesso a diretórios e arquivos não ocorreram. Dessa forma, são objetos futuros de estudo desse trabalho um estudo mais aprofundado do funcionamento de ACL no HDFS e o seu funcionamento com o Hue para realizar o controle de acesso a múltiplos arquivos e usuários de forma satisfatória.

Referências

Bertino, Elisa. 2015. "Big Data - Security and Privacy." In 2015 IEEE International Congress on Big Data, IEEE, 757–61.

Hadoop. "The Apache Hadoop." http://hadoop.apache.org/.

HadoopIssues. "Hadoop Issues Tracking." https://issues.apache.org/jira/browse/HADOOP.

HDFS. "Hadoop Distributed File System." http://hadoop.apache.org/hdfs/.

HDFSIssues. "HDFS Issues Tracking." https://issues.apache.org/jira/browse/HDFS.

Hortonworks. "Hortonworks: Open Enterprise Hadoop." http://hortonworks.com.

Hue. "Hue - Hadoop User Experience - The Apache Hadoop UI." http://gethue.com/.

Shvachko, Konstantin, Hairong Kuang, Sanjay Radia, and Robert Chansler. 2010. "The Hadoop Distributed File System." In Proc. of the MSST - Symp. on Mass Storage Systems and Technologies, IEEE, 1–10.

Tabatabaei, Mahsa. 2014. "Evaluation of Security in Hadoop." KTH Royal Institute of Technology.

Tankard, Colin. 2012. "Big Data Security." Network Security 2012(7): 5–8.

Thusoo, Ashish, J.S. Sarma, Namit Jain, Zheng Shao, Prasad Chakka, Suresh Anthony, Hao Liu, Pete Wyckoff, and Raghotham Murthy. 2009. "Hive - A Warehousing Solution Over a Map-Reduce Framework." Proceedings of the VLDB Endowment 2(2): 1626–29.

White, Tom. 2012. Hadoop: The Definitive Guide, 3rd Edition. 3rd ed. O'Reilly Media.

Zikopoulos, Paul C., Chris Eaton, Dirk DeRoos, Thomas Deutsch, and George Lapis. 2012. Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. McGraw-Hill.