ESTUDO DE FERRAMENTAS DE BIG DATA PARA DEFINIÇÃO DE INDICADORES EM TOMADA DE DECISÃO

NEVES, Lincon[1]
JARDINI, Evandro de Araújo[2]

RESUMO

O volume de dados que chega às organizações por meio do uso da internet está em constante crescimento. Atualmente, esses dados são coletados em canais, que variam desde cliques em propagandas, até as redes sociais, promovendo uma variedade muito maior de informação disponível. As tecnologias criadas para analisar e extrair informações desses dados são conhecidas como tecnologias de *big data* e, normalmente, possuem o objetivo de auxiliar nas tomadas de decisões de uma organização, seja reduzindo custos ou encontrando tendências e preferências para produtos e serviços por meio da mineração de dados. O big data exige novas formas de processamento e armazenamento para o seu real proveito, resultando, assim, na criação de diversas aplicações voltadas à manipulação de grandes coleções de dados. O presente trabalho visa a analisar ferramentas de processamento de big data objetivando estudos sobre os resultados apresentados por elas, encontrando indicadores úteis para as tomadas de decisões e recomendações. Para isso, foram estudadas as ferramentas Hadoop e Big data e desenvolveu-se uma aplicação capaz de gerar recomendações de filmes, livros e músicas de acordo com o histórico de opiniões de usuários armazenados em banco de dados.

Palavras-chave: Big data. Mineração de dados. Tomada de decisão. Big data. Hadoop.

[1] Graduando em Análise e Desenvolvimento de Sistemas pelo Instituto Federal de Educação, Ciência e Tecnologia de São Paulo - Campus Votuporanga. E-mail: linconneves@outlook.com

[2] Possui graduação em Tecnologia Em Processamento de Dados pelo Centro Estadual de Educação Tecnológica Paula Souza (1996), mestrado em Ciências da Computação pela Universidade de São Paulo (2000) e doutorado em Engenharia Elétrica pela Universidade de São Paulo (2007). Atualmente é professor com dedicação exclusiva do Instituto Federal de Educação, Ciência e Tecnologia de São Paulo campus Votuporanga e coordenador de área de informática. E-mail: eajardini@ifsp.edu.br

ABSTRACT

The volume of data that comes to organizations through the use of the Internet is constantly growing. Currently these data are collected in channels ranging from clicks on advertisement, to social networks, promoting a bigger variety of available information. The technologies designed to analyze and extract information from these data are known as big data technologies, and usually have the objective of assisting in decision-making in an organization, either by reducing costs or finding trends and preferences for products and services through data mining. The big data require new ways of processing and storage to its better use, resulting in the creation of various applications related to handle large collections of data. The present work aims to analyze these tools focusing on the development of a software capable of performing analyzes such as data mining, finding useful indicators for decision making. The present work aims at to analyze big data processing tools aiming at studies about the results presented by them, finding useful indicators for decision making and recommendations. To that end, the Hadoop and Big data tools were studied and an application was developed capable of generating recommendations of films, books and music according to the opinions history of users stored in database.

Keywords: Big data. Data mining. Decision making. Big data. Hadoop.

INTRODUÇÃO

Big data é o termo utilizado para nomear uma coleção de dados tão grande que está além da capacidade de aplicativos comumente utilizados para gerenciamento e análise de dados, requerendo novas formas de processamento para permitir a tomada de decisão aprimorada, a descoberta de conhecimento e otimização de processos (ALVES, 2015).

A tomada de decisão é complexa e deve levar em conta todas as alternativas, assim como prever suas consequências. Para que isso ocorra, é necessário obter a maior quantidade de indicadores úteis disponíveis. Assim, o big data torna-se uma grande oportunidade para empresas e consumidores. Marcelo Kekligian (HARVARD BUSINESS REVIEW**,** 2013) afirma que o big data

oferece a oportunidade de obter uma compreensão mais profunda das atitudes, preferências e comportamentos de seus clientes, tornando cada interação mais relevante, oportuna, segura e rentável. Os consumidores ganham a oportunidade de receber maior valor de seus bancos, seus fornecedores e outras empresas por meio de serviços mais rápidos, relevantes e personalizados.

Para extrair valores do big data que poderão servir de indicadores, uma das técnicas utilizadas é a mineração de dados. Han, Kamber e Pei (2012) descrevem a mineração de dados como um processo no qual métodos inteligentes são aplicados para extrair padrões de dados.

Bill Sweeney (HARVARD BUSINESS REVIEW, 2013), fundador da Risk Data and Analytics, assegura que o big data pode ser usado como uma forma de análise de riscos na qual as empresas têm potencial para melhor identificar os perigos não explícitos e fazer uma observação mais acurada de suas causas, aproveitando metadados e usando segmentação de clientes para identificar fatores de risco. É importante, contudo, destacar que o big data não é uma tecnologia específica para a área empresarial, seu conceito está migrando para todos os campos do conhecimento humano, pois, em essência, seu avanço é uma continuação da antiga busca da humanidade em medir, registrar e analisar o mundo (SCHÖNBERGER-MAYER; CUKIER, 2013).

1 CONTEXTO

Segundo dados da HP (2015), há um problema de análise de dados em empresas. 87% delas sãos incapazes de explorar totalmente seus dados e 85% são incapazes de analisar os dados rápido o bastante, desperdiçando, com isso, recursos em sua coleta, armazenamento e processamento. Há diversas soluções disponíveis no mercado para que as empresas possam trabalhar com a mineração de dados e big data. No entanto, o custo da licença de uma plataforma capaz de processar grandes quantidades de informações que não podem ser facilmente interpretadas manualmente pode ser bastante elevado (TREE INTELLIGENCE, 2014). Além disso, trabalhar com grupos de consultoria pode ser muito caro, lento e insuficiente para empresas de médio porte. Se a tomada de decisões, por meio de dados, torna-se regra em uma organização, as

informações devem estar disponíveis sem intermediários e de forma útil (AIELO, 2015).

Portanto, faz-se necessário estudar as ferramentas gratuitas na tentativa de substituir esses serviços pagos, minimizando os gastos para se trabalhar com essa tecnologia e tornando-a popular e adequada ao contexto de uso da empresa, o que pode resultar em tomadas de decisões mais precisas.

2 FUNDAMENTAÇÃO TEÓRICA

Esta seção apresenta um resumo dos conceitos e tecnologias relacionados com as pesquisas e desenvolvimentos realizados envolvendo: os cinco “Vs” do big data, o framework Hadoop e o framework Mahout.

2.1 Cinco “Vs” do big data

A seguir, serão detalhados os cinco “Vs do big data”, uma vez que as aplicações criadas para lidar com o *big data* precisam focar, no mínimo, em uma delas.

- Volume: característica mais simples do big data, pois não se está lidando com essa área se o volume dos dados não for realmente grande. Não existe uma regra que defina o tamanho desse volume, mas, normalmente, se trata da ordem de *gigabytes* de conteúdo.
- Velocidade: é fundamental que os dados sejam processados de forma rápida, pois muitos sistemas trabalham a fim de que se obtenham respostas em tempo real.
- Variedade: A variedade engloba os diferentes tipos de dados que podem estar em um big data. Além de textos e valores, os sistemas trabalham, hoje em dia, com diversos tipos de mídia, como vídeos, imagens, áudios, dentre outros.
- Veracidade: com relação à mineração de dados, a veracidade diz respeito ao quão verdadeiro é o indicativo retornado. As decisões que serão tomadas com base nesses valores

encontrados necessitam que eles sejam o mais próximos possível da realidade.

- Valor: os indicativos retornados pela mineração, por mais reais que sejam, devem subsidiar as melhores tomadas de decisão, por meio das quais se terá o retorno por todo o investimento em um sistema de big data.

2.2 Apache Hadoop

O Hadoop permite o processamento distribuído de grandes conjuntos de dados por meio de clusters de computadores, projetado para garantir larga escalabilidade partindo de um único servidor até um cluster com milhares de máquinas, cada uma oferecendo capacidade de computação e armazenamento local (FONTE, 2013). Ele trabalha principalmente com três das cinco características do big data*:* o volume, a velocidade e a variedade. Por meio da clusterização, permite que o processamento de grandes volumes de dados seja feito de forma rápida. Levando em consideração também que o Hadoop trabalha independentemente do tipo de dado processado, garante-se o fator variedade.

2.2.1 MapReduce

O MapReduce é um framework que permite realizar processamento em paralelo em um cluster Hadoop. Segundo Alves (2015), o MapReduce consiste em um JobTracker executado no nó mestre e responsável por gerenciar a execução dos serviços, distribuindo o trabalho para os nós TaskTrackers. Os TaskTrackers criam um processo separado para cada tarefa a fim de se certificar de que uma falha no processo não resulte em uma falha de TaskTracker. Eles ainda se reportam ao JobTracker em intervalos regulares, especificando que estão funcionais e trabalhando. Caso contrário, o nó é considerado morto e seu trabalho é enviado para um outro TaskTracker disponível.

2.2.2 Hadoop Distributed File System

Também conhecido pela sigla HDFS, é um sistema de arquivos distribuído, projetado para armazenar arquivos muito grandes e que não deve ser utilizado para aplicações que precisem de acesso rápido a um arquivo, e sim para

aplicações nas quais é necessário ler uma quantidade muito grande de dados (FONTE, 2013).

Funciona em uma estrutura necessitando de dois tipos de nós de armazenamento: um Namenode (mestre) e um ou mais Datanodes (servos). O Namenode controla todo o sistema de arquivos, mantendo metadados para todos os arquivos e diretórios da árvore de diretórios e arquivos do sistema. Já os Datanodes armazenam os blocos e enviam relatórios ao Namenode, periodicamente, com as listas dos blocos que eles estão armazenando. Focado em garantir alta confiabilidade e disponibilidade, o HDFS possui a replicação dos dados em dois nós: uma no próprio nó e outra em um diferente.

A Figura 1 ilustra a arquitetura do Hadoop:

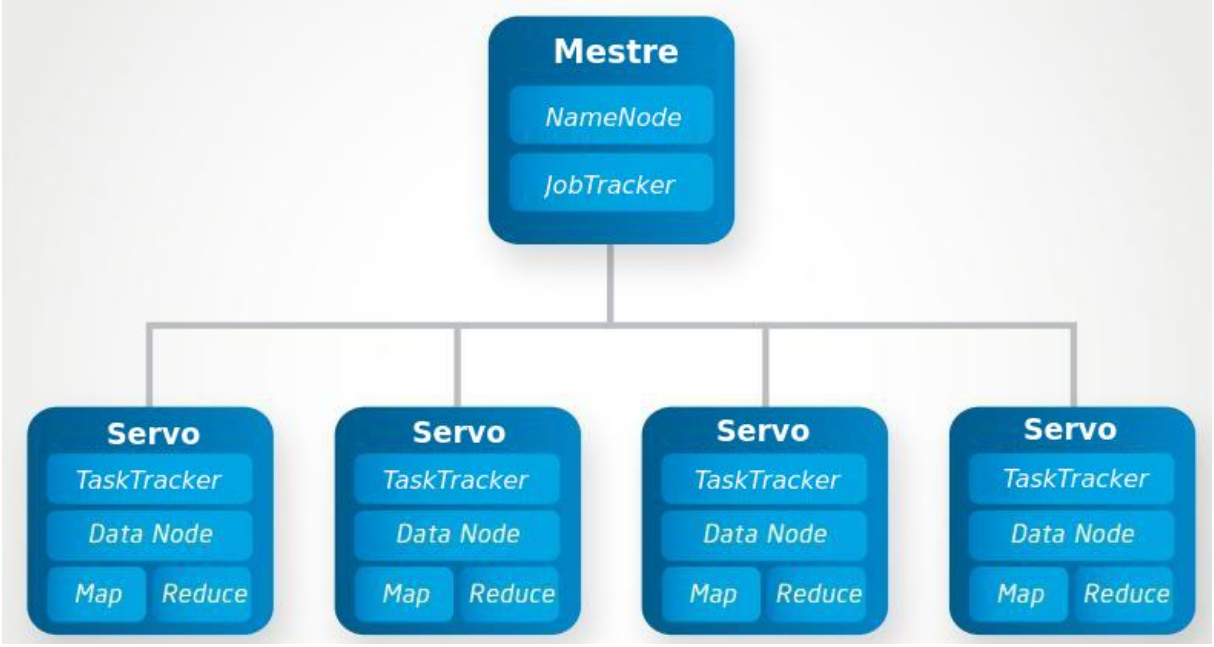

Figura 1 - Arquitetura de um cluster Hadoop.

Fonte: Adaptado de Pina, 2016.

2.3 Knowledge Discovery from Data (KDD)

Costuma-se tratar a mineração de dados como expressão de sentido semelhante ao termo em inglês *knowledge discovery from data*, enquanto outros veem a mineração de dados meramente como um passo essencial no processo de descoberta do conhecimento (HAN, KAMBER e PEI, 2012).

Sendo assim, a mineração de dados é a etapa mais importante na extração de conhecimento. Para Disner (2014), é nessa etapa que os dados são, efetivamente, transformados em conhecimento por meio de algoritmos que os exploram em busca de associações entre variáveis e estabelecendo relações entre elas, revelando padrões que podem ser importantes na interpretação desses dados.

2.4 Apache Mahout

Apache Mahout, assim como o MapReduce, é um framework que possui uma compreensiva lista de algoritmos focados em classificação, agrupamento (clustering) e recomendação. A biblioteca é focada em prover escalabilidade trabalhando sobre o Apache Hadoop para prover computação e armazenamento distribuídos. Seu conjunto de classes implementam algoritmos de mineração de dados e classes auxiliares. Foi desenvolvido utilizando recursos da API fornecida pelo Apache Hadoop, de modo a fazer uso do sistema de arquivos distribuídos, HDFS, e do framework MapReduce. Também é possível utilizar o Apache Mahout por meio de um aplicativo em linha de comando. Trata-se de um script, que faz uso de um aplicativo desenvolvido com base na API da ferramenta, disponibilizando algoritmos de mineração e funcionalidades de transformação e visualização de dados. Para utilizá-lo, informam-se a funcionalidade desejada e os parâmetros necessários para o seu funcionamento (PEREIRA, 2010).

2.5 Sistemas de Recomendação

Com o crescimento do e-commerce, hoje os consumidores possuem vários tipos de bens e serviços disponíveis de forma *online*. Os sistemas de recomendação auxiliam os consumidores fazendo recomendações de produtos que podem ser de interesse para o usuário, como livros, filmes, notícias e outros serviços (HAN, KAMBER e PEI, 2012).

Esse tipo de sistema pode usar uma abordagem baseada em conteúdo, uma abordagem colaborativa ou uma abordagem híbrida (que combine ambos os métodos). A abordagem baseada em conteúdo recomenda itens semelhantes àqueles que o usuário preferiu ou consultou no passado. A abordagem colaborativa recomenda itens com base nas opiniões de outros clientes que

possuem gostos ou preferências semelhantes aos do usuário, buscando semelhanças entre itens e preferências do cliente

3 DESENVOLVIMENTO

A metodologia empregada consistiu na pesquisa bibliográfica sobre o tema, no levantamento de ferramentas da área de mineração de dados e big data e, por fim, na realização de testes com as ferramentas.

Primeiramente, foi elaborada uma lista das tecnologias comumente empregadas em big data como a plataforma R, o software Weka, entre outras. Seguindo o princípio dos cinco “Vs” do big data, foram realizados estudos introdutórios com essas tecnologias e, em seguida, foram definidas os frameworks Hadoop e Mahout para serem estudadas em profundidade.

O segundo passo foi a implementação das tecnologias elencadas no primeiro passo, focando a biblioteca de recomendações. Foram submetidos conjuntos de dados obtidos na web contendo diversas opiniões de usuários de *sites* como MovieLens e Amazon e formado um *ranking* de recomendações de acordo com a preferência do usuário daqueles *sites*. Ambas as tecnologias foram unidas por meio de uma aplicação gráfica, desenvolvida neste trabalho, capaz de receber um conjunto de dados contendo avaliações de usuários a respeito de um tema, como filme, música, dentre outros, o que gera, dessa forma, recomendações para um determinado usuário baseadas nas notas que ele atribuiu a diversos outros itens pertencentes àquele conjunto de dados. A utilização dessas tecnologias possibilitou que a aplicação desenvolvida fosse capaz de gerar recomendações fazendo uso de cluster Hadoop, permitindo trabalhar com grandes volumes de dados.

3.1 Geração das Recomendações

É importante que os dados de entrada estejam previamente formatados em uma estrutura de, pelo menos, três colunas: a primeira identificando um usuário; a segunda identificando um item; e a última identificando o grau de preferência do usuário para aquele item, como demonstrado na Figura 2.

```
1.    1     1     5
2.    1     2     3
3.    1     3     4
4.    1     4     3
5.    1     5     3
6.    1     6     5
7.    1     7     4
8.    2     2     1
9.    2     3     5
10.   2     4     3
11.   2     6     2
12.   2     7     5
13.   2     8     5
14.   3     1     5
15.   3     4     5
16.   3     5     5
17.   3     6     3
18.   3     7     4
19.   3     8     5
```

Figura 2 - Trecho de um documento CSV com dados de entrada.
Fonte: O Autor, 2016.

No exemplo da Figura 3, foi utilizado um dataset disponibilizado pelo *site* MovieLens (movielens.org). Nesse dataset, encontram-se as avaliações que os usuários atribuem para os filmes vistos por eles. O objetivo desse código é buscar os vinte filmes que ainda não foram vistos pelo usuário 1 (user_id=1) e que sejam os mais indicados a ele baseado em seus gostos. Primeiro, cria-se um DataModel. Em seguida, deve-se preenchê-lo por meio dos dados armazenados em um banco PostgreSQL, na tabela “usuario_preferencia”, como demonstrado na linha 14 do quadro. Após isso, cria-se uma correlação entre os usuários por meio do UserSimilarity para, então, agrupá-los no próximo método, UserNeighborhood (linhas 17 e 19, respectivamente). Por fim, todos os atributos criados são repassados para uma implementação da classe Recommender (linha 22). O método recommend dessa classe é responsável por informar os filmes mais indicados para o usuário informado. Na linha 25, foram requisitados os vinte filmes mais indicados para o usuário identificado pelo número 1. Na linha 27, encontra-se uma estrutura de repetição para informar ao usuário as recomendações encontradas, as quais podem ser vistas na figura 4. As recomendações, além de possuir o número de identificação do item, possuem uma estimativa de como o usuário avaliaria aquele item.

```
1.    import org.apache.mahout.cf.taste.common.TasteException;
2.    import org.apache.mahout.cf.taste.impl.model.file.FileDataModel;
3.    import org.apache.mahout.cf.taste.impl.neighborhood.NearestNUserNeighborhood;
4.    import org.apache.mahout.cf.taste.impl.recommender.GenericUserBasedRecommender;
5.    import org.apache.mahout.cf.taste.impl.similarity.PearsonCorrelationSimilarity;
6.    import org.apache.mahout.cf.taste.model.DataModel;
7.    import org.apache.mahout.cf.taste.neighborhood.UserNeighborhood;
8.    import org.apache.mahout.cf.taste.recommender.RecommendedItem;
9.    import org.apache.mahout.cf.taste.recommender.Recommender;
10.   import org.apache.mahout.cf.taste.similarity.UserSimilarity;
11.
12.   DataModel model;
13.
14.   model = new PostgreSQLJDBCDataModel(dataSource, "usuario_preferencia",
15.   "user_id", "item_id", "preference", null);
16.
17.   UserSimilarity similarity = new PearsonCorrelationSimilarity(model);
18.
19.   UserNeighborhood neighborhood = new NearestNUserNeighborhood(100,
20.    similarity, model);
21.
22.   Recommender recommender = new GenericUserBasedRecommender(model,
23.    neighborhood, similarity);
24.
25.   List<RecommendedItem> recommendations = recommender.recommend(1, 20);
26.
27.   for (RecommendedItem recommendation : recommendations) {
28.                 System.out.println(recommendation);
29.   }
```

Figura 3 - Exemplo de método para gerar recomendações para um usuário.

Fonte: O Autor, 2016

Dessa maneira, considerando o resultado demonstrado na Figura 4, notam-se as seguintes recomendações feitas pelo algoritmo ao usuário: os itens que mais são recomendados são os números 1005, 1277, 482, 604, 640, 427, 705, 511 e 478, cujo valor da recomendação é 5. Em seguida, vêm os itens 283, 902 e assim sucessivamente, de acordo com o valor da recomendação, que varia de 0 a 5, sendo 0 para o menos recomendado e 5 para o mais recomendado.

```
RecommendedItem[item:1005, value:5.0]
RecommendedItem[item:1277, value:5.0]
RecommendedItem[item:482, value:5.0]
RecommendedItem[item:604, value:5.0]
RecommendedItem[item:640, value:5.0]
RecommendedItem[item:427, value:5.0]
RecommendedItem[item:705, value:5.0]
RecommendedItem[item:511, value:5.0]
RecommendedItem[item:478, value:5.0]
RecommendedItem[item:283, value:4.787457]
RecommendedItem[item:902, value:4.77585]
RecommendedItem[item:603, value:4.771144]
RecommendedItem[item:520, value:4.745629]
```

Figura 4 - Itens recomendados para o usuário informado.

Fonte: O Autor, 2016.

Para a utilização dos testes do algoritmo Recommender foi desenvolvida uma aplicação gráfica na linguagem Java que facilitasse o uso das tecnologias envolvidas. Dessa maneira, dentro da aplicação fez-se a importação da biblioteca Mahout como mostram as linhas 1 a 10 da figura 3. A execução do programa fazendo uso do cluster Hadoop dá-se pela linha de comando:

#hadoop jar programa_java_executável

No parâmetro *programa_java_executável,* insere-se o caminho para a aplicação*;* a execução do Mahout sobre o Hadoop é feita de forma transparente.

A Figura 5 demonstra a interface criada para a geração de recomendações.

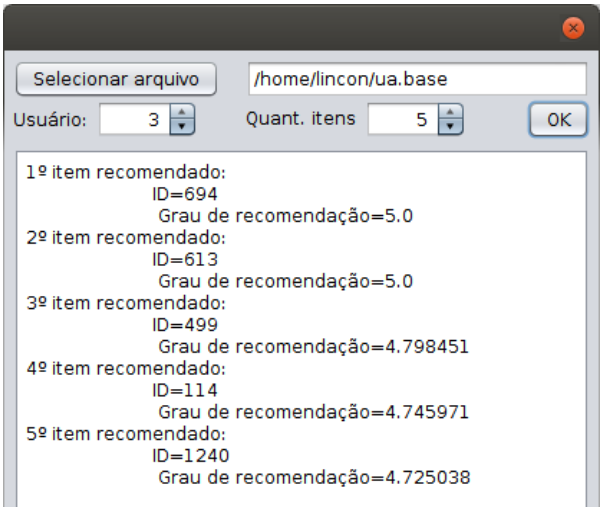

Figura 5 - Interface criada para a recomendação de itens.
Fonte: O Autor, 2016.

Por meio do botão “Selecionar arquivo”, seleciona-se o arquivo de texto obtido na *web* que contém os dados de entrada formatados na estrutura mencionada anteriormente. Após isso, no campo “Usuário”, informa-se o identificador do usuário para o qual se deseja recomendar os itens, e em “Quant. itens”, a quantidade de itens recomendados. No espaço abaixo desses campos, ao clicar em “OK”, será gerada uma lista de itens recomendados ordenados pela ordem de mais indicado para o menos indicado.

4 RESULTADOS

Como resultados, são demonstradas mais duas recomendações, sendo uma para livros e outra para produtos alimentícios. Para cada recomendação é utilizado um conjunto de dados (dataset) representado por um arquivo, contendo as classificações dos usuários atribuídas aos respectivos temas. Seguem as descrições de cada dataset:

- Book-Crossing Dataset: conjunto de dados obtidos no site http://www2.informatik. uni-freiburg.de/~cziegler/BX/, que contém as classificações de usuários a respeito de livros. Para cada classificação é atribuída uma nota de 0 a 10, sendo 0 para “não gostou” e 10 para “gostou muito”. O dataset contém 278.858 usuários (anônimos, mas com informações demográficas), fornecendo 1.149.780 avaliações de cerca de 271.379 livros.
- Ratings Grocery and Gourmet Food Dataset: conjunto de dados obtidos no site http://jmcauley.ucsd.edu/data/amazon/, o qual contém as classificações de usuários a respeito de alimentos. Para cada classificação é atribuída uma nota de 0 a 5, sendo 0 para “não gostou” e 5 para “gostou muito”. O dataset contém 1.297.156 avaliações de produtos alimentícios vendidos pelo site da Amazon.

A demonstração das recomendações para o conjunto de dados de livros pode ser observada na Figura 6.

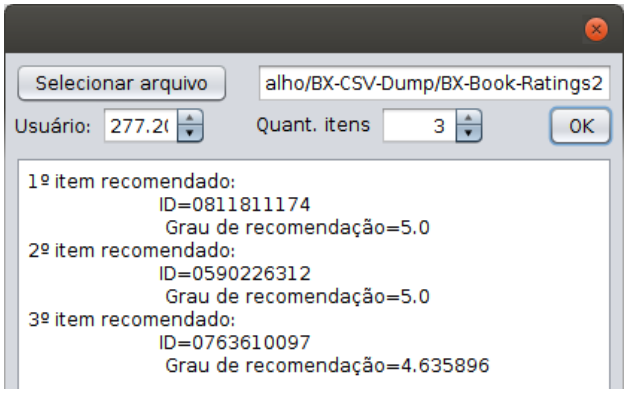

Figura 6 - Geração de uma lista de livros recomendados.

Fonte: O Autor, 2016.

Utilizando o ID do livro informado na lista gerada, que corresponde ao ISBN da obra, pode-se procurar os livros na web. Constatou-se que todos os livros são do gênero infantil, o que fornece uma certa garantia de que os itens recomendados estão, de certa forma, relacionados. Os livros em questão são Boomer Goes to School, e Don`t You Feel Well, Sam?

No segundo dataset, foi necessário realizar uma formatação do arquivo para que ele fosse corretamente lido pela aplicação. Os IDs dos usuários e produtos eram extensos e utilizavam letras em conjunto com números para formar o identificador, o que impedia a aplicação de gerar a recomendação, pois ela apenas trabalha com números. Uma forma simples de resolver esse problema é atribuir um novo identificador único a cada usuário e produto. A Figura 7 mostra a recomendação gerada por esse dataset.

Figura 7 - Geração de uma lista de alimentos recomendados.
Fonte: O Autor, 2016.

Ao se verificar no *site* da Amazon os produtos recomendados, mais uma vez pode-se notar uma certa relação entre os itens, considerando que ambos pertencem a uma categoria de chás japoneses.

CONSIDERAÇÕES FINAIS

Este é um trabalho introdutório e há muito o que ser estudado, pois a tecnologia de big data em união com o Mahout possui diversos recursos que não foram abordados. O que se fez foi conhecer o assunto, realizar testes com bases de dados prontos disponibilizadas na internet para esse fim e recomenda-se um estudo utilizando outros algoritmos dentro do Mahout, assim como a

implementação do modo gráfico para facilitar seu uso por qualquer pessoa, e não apenas aquelas com domínio em informática.

Os sistemas de recomendação são uma ótima maneira de demonstrar as ferramentas, pois aborda uma utilização real, presente em várias organizações e, principalmente, em e-commerces, abrangendo tanto o big data como também a mineração de dados. Por utilizar uma mesma base de entrada, ou seja, avaliações dos próprios usuários, os sistemas de recomendação não se limitam a recomendar apenas produtos, mas podem vir a contribuir, gerando recomendações de outros conteúdos, como notícias em *sites* e, até mesmo, na medicina de precisão, por meio do monitoramento constante de pacientes e tratamentos médicos. Para trabalhos futuros, sugere-se utilizar datasets que usem outros dados, além de avaliações, como histórico de cliques.

Outro ponto percebido é que não foi necessário obter o conhecimento da identidade dos usuários para recomendar-lhes itens, o que não causa preocupação daqueles que primam pelo sigilo de seus dados pessoais.

Conclui-se que há diversas ferramentas gratuitas como o Hadoop e o Mahout disponíveis para se trabalhar sem que seja necessária a contratação de serviços terceirizados. Porém, é necessário afirmar que tanto o big data quanto a mineração de dados são áreas sempre em desenvolvimento e muito difíceis de se lidar. As organizações interessadas em trabalhar nesse meio devem ter funcionários preparados, caso contrário, há uma grande chance de se gastar recursos de forma equivocada.

REFERÊNCIAS

AIELO, Rafael. **Big data não se aplica apenas a grandes empresas.** 2015. Disponível em: <http://convergecom.com.br/tiinside/services/15/09/2015/big-data-nao-se-aplica-apenas-a-grandes-empresas/>. Acesso em 31 out. 2016.

ALVES, Atos Ramos. **Infraestrutura Big data com OpenSource**. Rio de Janeiro: Ciência Moderna, 2015.

DISNER, Daniel da Silva. **Mineração de dados para obtenção de conhecimento em Big data.** 2014. 41 f. TCC (Graduação) - Curso de Sistemas de Informação, Centro Universitário Eurípides de Marília, Marília, 2014.

FONTE, Flávio. **O que é o Hadoop?** 2013. Disponível em: <http://bigdatabrazil.blogspot.com.br/2013/06/o-que-e-o-hadoop.html>. Acesso em: 10 nov. 2015.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. **Data Mining:** concepts and techniques. 3. ed. Waltham: Morgan Kaufmann, 2012.

HARVARD BUSINESS REVIEW. Big data: the future of information and business. 2013. Disponível em: <https://hbr.org/resources/pdfs/comm /experian/hbr_serasa_experian_report.pdf>. Acesso em: 07 nov. 2015.

HP. **Soluções de Big data.** 2015. Disponível em <http://www8.hp.com/br/pt/business-solutions/big-data.html>. Acesso em 31 out. 2016.

MAYER-SCHÖNBERGER, Viktor; CUKIER, Kenneth. **Big data**: como extrair volume, variedade, velocidade e valor da avalanche de informação cotidiana. Rio de Janeiro: Elsevier Editora Ltda., 2013.

PEREIRA, Adriano. **Mineração de dados distribuída e escalável usando Apache Mahout.** 2010. 59 f. TCC (Graduação) - Curso de Ciência da Computação, Universidade Federal de Santa Maria, Santa Maria, 2010.

PINA, Antonio Carlos. **Introdução ao BigData para IT pros:** parte 2. 2016. Disponível em: <https://www.linkedin.com/pulse/introdução-ao-bigdata-para-pros-parte-2-antonio-carlos-pina>. Acesso em: 31 jun. 2016.

TREE INTELLIGENCE. Big data: vale a pena criar uma equipe? 2014. Disponível em <http://treeintelligence.com/big-data-vale-a-pena-criar-uma-equipe/> Acesso em 31 out. 2016.