Utilizando um cluster virtual com Hadoop como uma ferramenta para exploração de big data em processamento de imagens digitais

Marcelo Musci [1]
Patrick Nigri Happ [1]
Gilson Alexandre Ostwald Pedro da Costa [2]
Raul Queiroz Feitosa [1,2]

[1] Pontifícia Universidade Católica do Rio de Janeiro – PUC-Rio
Caixa Postal 38097 - 22453-900 – Rio de Janeiro - RJ, Brasil
{musci, patrick, raul}@ele.puc-rio.br

[2] Universidade Estadual do Rio de Janeiro – UERJ
Rua São Francisco Xavier, 524 – 5º andar - 20550-900 - Maracanã - RJ, Brasil
gilson.costa@ime.uerj.br

Abstract The amount of available remote sensing (RS) data is increasing at an extremely rapid pace due to recent advances in Earth observation technologies. This scenario leads to new challenges related to the ability to handle huge volumes of data with respect to computational techniques and resources. In this sense, RS data processing can be considered a *big data* problem, and in this context cloud computing is a trend since it offers a powerful infrastructure to perform large-scale computing, which is usually available in a *pay-as-you-go* model, and alleviates users of the need to acquire and maintain a complex computing infrastructure. Although prices currently practiced by cloud infrastructure providers are reasonably low, the development and testing of cloud-based platforms is a long work, which may become unfeasible considering the total costs involved. This work describes a solution to the problem of the costs involved in the development of methods based on cloud computing, in particular for RS data processing tools based on the Hadoop framework. Such a solution is based on the creation of a configurable virtual cluster on a single physical machine, installed with the software components required to run a distributed application. The virtual infrastructure provided by the solution was used for the development and testing of extensions of a recently proposed architecture for the distributed classification of RS data. To validate the extensions, classification experiments were carried out on hyperspectral images acquired with the ROSIS sensor, covering the University of Pavia in Italy.

Palavras-chave: computação em nuvem; Hadoop; cluster; big-data.

1. Introdução

A quantidade de dados de sensoriamento remoto (SR) disponíveis está aumentando a um ritmo extremamente rápido devido aos recentes avanços nas tecnologias de Observação da Terra (Ullah et al., 2015; Datcu, 2015; Zhang, 2015). Centenas de satélites estão presentemente em órbita, adquirindo grandes quantidades de informação sobre a superfície da Terra todos os dias. As melhorias relacionadas com a resolução espacial, a frequência de revisita e o número de bandas espectrais são os principais impulsionadores dessa crescente disponibilidade de dados. Por exemplo, o Sentinel-1, da Agência Espacial Europeia, gera sozinho cerca de 1,5 GB por dia (Grabak, 2014), e o projeto EOSDIS da NASA produz cerca de 16 TB de dados por dia (NASA, 2015). Este cenário leva a novos desafios, relacionados com a capacidade de lidar com enormes volumes de dados (Kishor, 2013; Lee e Kang, 2015), no que diz respeito a técnicas computacionais e recursos.

Nesse sentido, o processamento de dados de SR pode ser considerado um problema de *big data*, devido ao alto volume de dados (TB / dia), à variedade destes dados (imagens óticas, de radar, hiperespectrais, etc.) e à velocidade de geração dos dados a serem processados (Ma et al., 2015; Schade, 2015). Felizmente, lidar com grandes dados é atualmente um problema comum enfrentado por diferentes áreas na indústria e centros de pesquisa. Nesse contexto, a computação em nuvem é uma tendência (Fernandez , 2014), uma vez que oferece uma

poderosa infraestrutura para executar computação em grande escala, que geralmente está disponível em um modelo *pay-as-you-go*, e que alivia os usuários da necessidade de adquirir e manter uma complexa infraestrutura computacional.

É crescente o número de abordagens para o processamento de dados de SR em ambientes de computação em nuvem, como pode-se observar em alguns exemplos recentes: em (Ferreira et al., 2015) uma nova arquitetura para análise baseada em objetos foi apresentada; (Happ et al., 2016) introduz uma metodologia para a segmentação distribuída de imagens de SR; e (Ayma et al., 2016) descreve uma arquitetura para a classificação distribuída de dados de SR.
Os trabalhos mencionados no parágrafo anterior seguem a tendência de utilizar para o processamento distribuído de grandes volumes de dados o paradigma de programação MapReduce (Dean e Ghemawat, 2004), mais especificamente a sua versão de código livre incluída no framework Hadoop (Hadoop, 2014), que fornece uma plataforma escalável, confiável e de baixo custo para processar e armazenar grandes quantidades de dados em clusters. Além disso, aqueles trabalhos foram implementados para serem executados em ambientes comerciais de infraestrutura em nuvem, como o AWS (Amazon Web Services). Outra característica daquelas abordagens é que elas são extensíveis, no sentido em que novas funcionalidades ou algoritmos podem ser incorporados às respectivas implementações.

Apesar dos preços atualmente praticados por provedores de infraestrutura em nuvem serem razoavelmente baixos, o desenvolvimento e teste de plataformas como as mencionadas é um trabalho longo, que pode se tornar inviável pelo total de custos envolvidos. Uma alternativa seria realizar as fases de desenvolvimento e testes sobre clusters virtuais de computadores, instalados em uma única máquina física. Existem atualmente algumas soluções prontas de clusters virtuais, como a Hortonworks HDP Sandbox (Hortonworks, 2016) e Cloudera QuickStart VM (Cloudera, 2016), porém estas soluções fornecem um único nó (host) que simula o paralelismo envolvido no Hadoop, não permitindo, por exemplo a aferição de *speedups*, mesmo que parciais, proporcionada pelos métodos em desenvolvimento.

Neste trabalho é descrita uma solução para o problema dos custos envolvidos no desenvolvimento de métodos baseados em computação em nuvem, em particular para ferramentas de processamento de dados de SR baseadas no *framework* Hadoop. Tal solução se baseia na criação de um cluster virtual configurável em uma única máquina física, instalado com os componentes de software necessários para a execução de uma aplicação distribuída.
A infraestrutura virtual fornecida pela solução foi utilizada para o desenvolvimento e testes de extensões da arquitetura proposta em (Ayma et al., 2016) para a classificação distribuída de dados de SR. Para validar as extensões foram realizados experimentos de classificação sobre imagens hiperespectrais do sensor óptico ROSIS sobre a Universidade de Pavia na Itália.

2. Arquitetura de Classificação

Em (Ayma et al., 2016) foi proposta uma arquitetura de computação em nuvem, projetada para permitir processos de classificação supervisionados em grandes volumes de dados de sensoriamento remoto. A arquitetura suporta a execução distribuída, comunicação de rede e tolerância a falhas capacidades de uma forma que é completamente transparente para o usuário. A arquitetura contém três camadas, que fornecem diferentes níveis de abstração e podem ser implementadas independentemente. Cada camada tem como alvo um tipo de usuário diferente, de acordo com sua especialidade. Além da descrição da arquitetura, outra importante contribuição de (Ayma et al., 2016) é que ele descreve como estender a arquitetura através da integração de diferentes algoritmos de classificação.

A arquitetura InterCloud Data Mining foi projetada para suportar a interação entre algoritmos de aprendizado de máquina e grandes conjuntos de dados, através da distribuição de dados e tarefas de processamento (classificação). A arquitetura contém três camadas de

abstração: a camada de definição de projetos; a camada de classificação; e a camada de distribuição.

A camada de definição de projetos permite a interação com usuário final. A informação fornecida pelo usuário através da camada de definição de projetos compreende todas as informações necessárias para a execução da aplicação de classificação, ou seja: o algoritmo de classificação a ser utilizado; os valores dos parâmetros de tal algoritmo; o número de nós de processamento a serem alocados no ambiente de computação em nuvem; a localização dos conjuntos de dados de treinamento e classificação (teste); entre outras.

Através da camada de classificação, os usuários com habilidades de programação convencional (em oposição aos usuários com habilidades de programação distribuída), são capazes de incorporar novos algoritmos de classificação em uma implementação da arquitetura. A camada de classificação é estruturada em uma linguagem de programação de alto nível que esconde a complexidade de lidar diretamente com o modelo de programação distribuída.

A implementação anterior do InterCloud Data Mining continha quatro algoritmos de classificação supervisionados do ambiente Weka: Näive Bayes, Decision Tree, Random Forest e Support Vector Machines (Waikato, 2014). Durante o desenvolvimento deste trabalho foram efetuados testes com os três últimos algoritmos no ambiente de desenvolvimento virtual proposto neste trabalho.

A linguagem de programação Pig Latin é a responsável pelo controle de todo processo de classificação, definindo os conjuntos de dados de treinamento e teste, selecionando o algoritmo de classificação e armazenando o resultado da classificação em um determinado repositório bem como fornece um método fácil de implementação para funções personalizadas definidas pelo usuário (user-defined functions – UDFs). O uso de Pig Latin torna mais fácil a tarefa de interagir com a camada de distribuição baseada em MapReduce (Dean and Ghemawat, 2008), sendo um dos modelos de programação mais populares para o processamento de grandes conjuntos de dados implementado no Apache Hadoop.

3. Criação do Cluster Virtual

A abordagem global consiste em criar uma máquina virtual com um software de virtualização como o VMWare Workstation (VMWare, 2016) ou o VirtualBox (VirtualBox, 2016) e efetuar as configurações necessários para atuar como um nó de cluster (especialmente as configurações de rede). Esta máquina virtual, com a distribuição GNU/Linux CentOS 7 instalada, é então clonada quantas vezes existirão os nós no cluster do Hadoop. Somente um conjunto limitado de mudanças torna-se necessário para finalizar o nó de forma a estar operacional, como a definição do nome do host e o endereço IP, além de ajustes diversos do Linux para manutenção da compatibilidade com o Hadoop.

Neste artigo foi criado um cluster Hadoop de quatro máquinas virtuais no ambiente VMWare. Para essa instalação foi utilizada a distribuição Hadoop da Hortonworks em conjunto com o assistente de instalação Apache Ambari. O Hortonworks Data Platform (HDP, 2016) é uma suíte de funcionalidades essenciais para implementação do Hadoop, que pode ser usado para qualquer plataforma tecnológica de dados e o Apache Ambari é um exemplo de console de gerenciamento do cluster Hadoop desenvolvido pelo fornecedor Hortonworks.

Nesse cluster foram instalados os serviços YARN, HDFS, MapReduce2, Pig, Hive, HBase, ZooKeeper e AmbariMetrics. Essencialmente somente os quatro primeiros são necessários, porém os outros serviços fornecem diversas funcionalidades importantes além de gerenciamento e análise de dados. O primeiro nó, que executará a maioria dos serviços de cluster, requer mais memória (16 GB) do que os outros 3 nós (8 GB). O tamanho do HD virtual utilizado para cada um dos quatro nós deve ser o maior possível, devido a replicação

dos dados no cluster proporcionada pela tolerância a falhas do Hadoop. No cluster construído foram reservados HDs de 80Gb para cada um dos nós.
Para os diversos experimentos foram habilitados sequencialmente os nós do cluster, começando com 1 (base), 2 e 4 nós de cada vez.

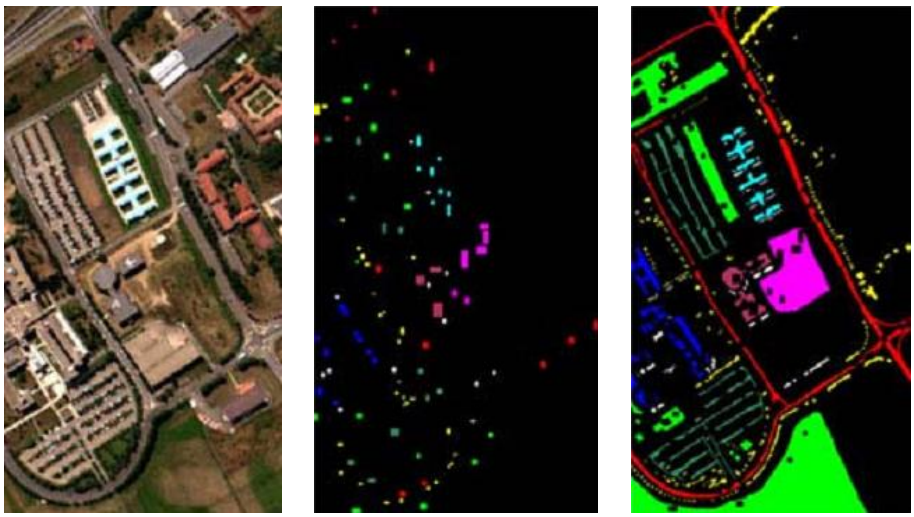
A máquina onde foi instalado o cluster virtual é um servidor Intel Xeon i7-3970X com 6 núcleos físicos (cores) e 12 lógicos, 3.5GHz e 64Gb de memória. As versões do Hadoop e Pig usadas foram 2.5 e 0.16 respectivamente

4. Resultados e Discussão

Esta seção apresenta os resultados da avaliação experimental da arquitetura proposta, realizando uma série de experimentos de classificação com os dados provenientes de uma imagem hiperespectral coletada pelo sensor óptico ROSIS sobre a Universidade de Pavia localizada na Itália, Figura 1(a). A imagem contem 610×340 pixels com 1.3-m de resolução especial e 103 bandas espectrais. As classes de interesse compreendem 21 tipos de cobertura do solo. A Figura 1(b) e (c) ilustram os conjuntos de pixels de treinamento e teste respectivamente. Somente nove componentes principais de cada pixel de referência foram utilizados para classificação, resultando em um conjunto de dados de 20Mb. A partir deste conjunto de dados de teste original foram construídos conjuntos de dados sintéticos replicando-o 20, 40, 200 e 300 vezes.

Os experimentos foram concebidos para avaliar a funcionalidade do cluster virtual para o desenvolvimento e testes de extensões da arquitetura proposta em (Ayma et al., 2016).

Nos testes realizados foram utilizados os algoritmos de classificação supervisionados do ambiente Weka: Decsion Trees, Random Forest e Support Vector Machines (SVM). Os parâmetros correspondentes aos algoritmos de classificação foram configurados de acordo com (Waikato, 2014), onde Random Forest possui 100 nós e uma semente aleatória fixa; o SVM possui classificação multiclasses, com núcleo polinomial, complexidade C = 1.0 e expoente γ=1.0 em uma validação cruzada com 5 *folds* e finalmente Decision Tree com poda C = 0.25 e instâncias M = 2.



(a) (b) (c)

Figura 1 – Imagem da base de dados Pavia: (a) Composição em falsa cor. (b) Pixels do conjunto de treinamento. (c) Pixels do conjunto de teste.

A Tabela 1 mostra o tempo de execução total da classificação (treino e teste) em segundos para a base de dados Pavia. A tabela indica que o tempo gasto na classificação diminui rapidamente de acordo com a adição de nós no cluster.

Nas Figuras 2 e 3 são apresentados os *speedup*s obtidos pelos classificadores Decision Tree e SVM respectivamente. As figuras mostram que a medida que o conjunto de dados aumenta a configuração do cluster produz melhores resultados. Volumes de dados maiores permitem *speedup*s mais altos, uma vez que os dados são distribuídos em mais nós.

Tabela 1.Tempo de execução em segundos da classificação da base de dados Pavia com o uso dos classificadores Decision Tree e SVM.

Classificador	**# Nós do Cluster**	**Tamanho do Arquivo de Dados**			
		444Mb	888Mb	4.2Gb	6.9Gb
Decision Tree	01 Nó	160s	270s	1175s	1973s
	02 Nós	130s	195s	990s	1583s
	04 Nós	110s	103s	344s	629s
SVM	01 Nó	140s	270s	1262s	2015s
	02 Nós	120s	203s	755s	1640s
	04 Nós	63s	104s	378s	669s

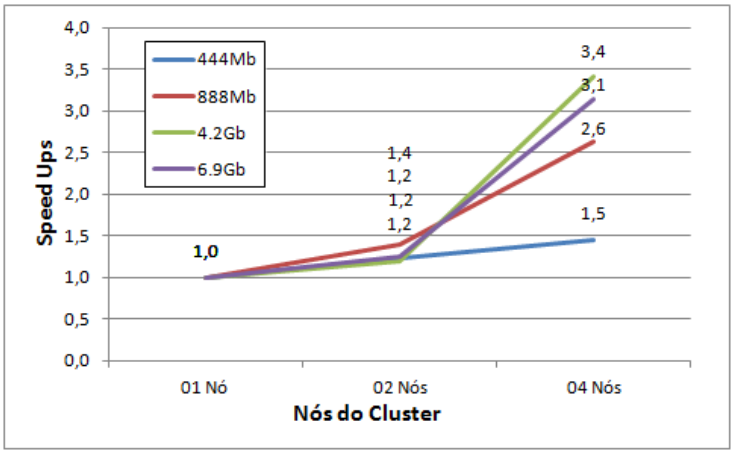


Figura 2 – Gráfico de *speedup* do classificador Decision Tree para a base de dados Pavia.

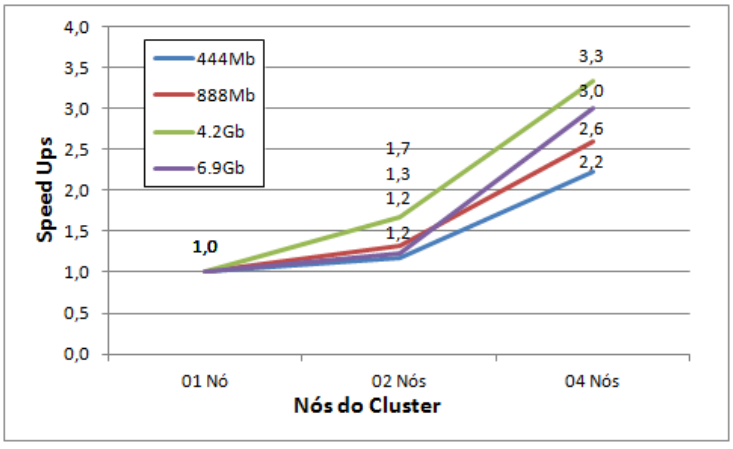


Figura 3 – Gráfico de *speedup* do classificador SVM para a base de dados Pavia.

5. Conclusões

Neste trabalho é descrita uma solução para o problema dos custos envolvidos no desenvolvimento de métodos baseados em computação em nuvem, em particular para ferramentas de processamento de dados de SR baseadas no *framework* Hadoop. Tal solução se baseia na criação de um cluster virtual configurável em uma única máquina física, instalado com os componentes de software necessários para a execução de uma aplicação distribuída.

Em (Ayma et al., 2016) foi proposta uma arquitetura de computação em nuvem, projetada para permitir processos de classificação supervisionados em grandes volumes de dados de sensoriamento remoto.

Neste trabalho foi proposta a utilização de um cluster virtual em uma única máquina física, com os componentes de software necessários para a execução de uma aplicação distribuída com Hadoop, como solução para o problema dos custos envolvidos no desenvolvimento de métodos baseados em computação em nuvem.

Uma implementação de um cluster virtual com quatro nós foi criado para validação. Essa implementação explora os benefícios de trabalhar em clusters com o *framework* Hadoop, fornecendo uma plataforma robusta e flexível que permite trabalhar com grandes conjuntos de dados em infraestruturas distribuídas.

A análise experimental, realizada com a utilização dos classificadores Decision Tree e SVM como funções personalizadas (UDFs) construídas inicialmente para a infraestrutura em nuvem utilizando Hadoop, apresentaram resultados satisfatórios no cluster virtual, descrito neste artigo, demonstrando a sua escalabilidade e o seu potencial para lidar com grandes conjuntos de dados.

À medida que o conjunto de dados aumenta, a adição de mais nós proporciona um processamento mais eficiente. Observou-se também que as velocidades aumentam com a quantidade de dados sendo processados. Isso ocorre porque, para conjuntos de dados maiores, os recursos distribuídos podem ser mais bem explorados, resultando em maior paralelização.

Trabalhos futuros envolvem a exploração de outras técnicas de processamento distribuído igualmente utilizadas em análise de grandes massas de dados como Tez e Spark.

Referências

Ayma, V.A.Q., Costa, G.A.O.P., Happ, P.N., Feitosa, R.Q., Ferreira, R.S., Oliveira, D.A.B. e Plaza, A. A New Cloud Computing Architecture for the Classification of Remote Sensing Data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, v. PP, p. 1-8, 2016.

Cloudera QuickStart VM Disponível em: <http://www.cloudera.com/downloads/quickstart_vms/5-8.html>, Acessado em: 10 de novembro de 2016.

Datcu, M. "HD-03: Big Data from earth observation: Analytics, mining and semantics," in IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Milan, 2015.

Dean, J,and Ghemawat, S. MapReduce: Simplified Data Processing on Large Clusters. Google Labs, OSDI; 2004 137–150.

Dean, J. and Ghemawat, S. “MapReduce: Simplified data processing on large clusters,” Commun. ACM, vol. 51, no. 1, pp. 107–113, Jan. 2008

Fernandez, A. et al., “Big data with cloud computing: An insight on the computing environment, mapreduce, and programming frameworks,” Wiley Interdisciplinary Rev., Data Mining Knowl. Disc., vol. 4, no. 5, pp. 380–409, Sep./Oct. 2014.

Ferreira, R. S. ; Oliveira, D.A.B. ; Happ, P. N. ; Costa, G.A.O.P. ; Feitosa, R. Q. ; Bentes, C. . InterIMAGE Cloud Platform: Em direção à arquitetura de uma plataforma distribuída e de código aberto para a interpretação automática de imagens baseada em conhecimento. In: XVII Simpósio Brasileiro de Sensoriamento Remoto -

SBSR, 2015, 2015, João Pessoa. Anais XVII Simpósio Brasileiro de Sensoriamento Remoto - INPE, 2015. p. 5264-5271.

Grabak, O. "Sentinel-1 Mission Status," in 15th Meeting of the International Ice Charting Working Group (IICWG), 2014.

Hadoop. Disponível em: <https://hadoop.apache.org>. Acesso em: 09.out.2014.

Happ, P.N., Costa, G. A. O. P., Bentes, C., Feitosa, R. Q., Ferreira, R. S. and Farias, R. "A cloud computing strategy for region-growing segmentation," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. PP, no. 99, pp. 1–10, 2016.

Hortonworks Data Platform (HDP), Apache Ambari Installation. Hortonworks. Disponível em: <http://docs.hortonworks.com/HDPDocuments/Ambari-2.4.1.0/bk_ambari-installation/content/ch_Installing_Ambari.html>, Acesso em: 10 de outubro de 2016.

Kishor, D. "Big Data: The New Challenges in Data Mining," International Journal of Innovative Research in Computer Science & Technology, vol. 1, no. 2, pp. 39-42, September 2013.

Lee, J. G. and Kang, M. "Geospatial Big Data: Challenges and Opportunities," Big Data Research, vol. 2, pp. 74-81, June 2015.

NASA EARTHDATA. (2015, February) EOSDIS Annual Metrics Reports. [Online]. https://earthdata.nasa.gov/about/system-performance/eosdis-annual-metrics-reports

Ma Y. et al., "Remote sensing big data computing: Challenges and opportunities," Future Generation Computer Systems, vol. 51, pp. 47-60, October 2015.

Schade, S. "Big data breaking barriers - first steps on a long trail," The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XL-7/W3, pp. 691-697, May 2015.

VirtualBox Disponível em: <https://www.virtualbox.org/>, Acesso em: 10 de outubro de 2016.

VMWare Workstation. Disponível em: <http://www.vmware.com/br/products/workstation.html>, Acesso em: 10 de outubro de 2016.

Ullah, M. et al., "Real-Time Big Data Analytical Architecture for Remote Sensing Application," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. PP, no. 99, pp. 1-12, May 2015.

Waikato University, Machine Learning Group. (2014) Weka 3: Data Mining Software in Java. [Online]. http://www.cs.waikato.ac.nz/~ml/weka/index.html

Zhang, L., Qian Du, and Datcu, M. "Special section guest editorial: Management and analytics of remotely sensed dig data," Journal of Applied Remote Sensing, vol. 9, no. 1, pp. 1-2, July 2015.