Arquiteturas de sistemas de visualização para grandes bases de dados da Observação da Terra por satélites

Matheus Monteiro Mariano

Laboratório de Computação e Matemática Aplicada, LAC
Instituto Nacional de Pesquisas Espaciais, INPE
São José dos Campos, SP - Brasil

matheus.mariano@inpe.br

Palavras-chave: Observação da Terra, Big Data, Cloud, Sensoriamento Remoto.

1. Introdução

A Observação da Terra envolve o conhecimento científico e tecnológico nos campos de sensoriamento remoto e geoprocessamento, levantamento de recursos naturais e monitoramento do meio ambiente, realizando atividades de pesquisa, desenvolvimento e aplicações nos campos de Sensoriamento Remoto e Processamento de Imagens Digitais. A detecção de objetos terrestres é uma técnica poderosa para fornecer informações rapidamente sobre os recursos da Terra em grandes extensões de áreas (Verhoef, 1985). Através desta capacidade uma nação consegue ter conhecimento sobre seu próprio território, sendo um importante fator que pode influenciar na tomada de decisões de projetos de médio e longo prazo de interesse da sociedade.

Com o passar do tempo, diversos satélites foram colocados em órbita da Terra para monitorar o planeta, com novas gerações de satélites continuam sendo criados e lançados por instituições acompanhando os avanços tecnológicos da sociedade para melhorar as imagens obtidas, aprimorando as resoluções espaciais, espectrais e temporais. Contudo, conforme as imagens se tornam mais detalhadas devido a capacidade de captação ser mais eficiente, mais pesada se torna o arquivo da imagem, e, com isso, os dados da Observação da Terra cada vez mais aumenta em tamanho e variedade (Nativi et al, 2015).

Desta forma, mesmo com imagens em alta qualidade é necessário ter métodos e heurísticas que possam interpretar essa massa de dados enviados pelos satélites de maneiras eficientes. Para isso, como os sistemas de visualizações de imagens orbitais não possuem opções de tratamento destes dados, diversos módulos de visualizações de grandes bases de dados de Observação da Terra foram criadas para tirar proveito da imensa quantidade de dados gerados pelos satélites. Estes módulos utilizam arquiteturas específicas para tratar estas grandes bases de dados a partir de conceitos de *Big Data*, um paradigma de análise de dados para bases heterogêneas, e NoSQL, focado na distribuição dos dados em diferentes locais utilizando uma rede (seja local ou internet).

Neste trabalho serão descritas algumas arquiteturas propostas pela literatura que podem ser empregadas em sistemas de visualizações de imagens orbitais. Como a área de *Big Data* foi altamente estudada durante os anos, serão destacadas três arquiteturas de diferentes épocas e finalidades, visando mostrar diferentes abordagens para tratar grandes bases de dados e como o conceito foi evoluindo na literatura. Este trabalho está dividido na seguinte forma: no

Capítulo 2 será dado uma descrição geral sobre grandes bases de dados, mostrando conceitos importantes de *Big Data*; no Capítulo 3 será introduzido as três arquiteturas para grandes bases de dados; no Capítulo 4 será feito uma crítica e a conclusão do trabalho; e por fim no Capítulo 5 será mostrado as referências.

2. Descrição Geral

2.1 Grandes bases de dados de Sensoriamento Remoto (SR)

O Sensoriamento Remoto (SR) é um conjunto de tecnologias de medição de características de um objeto ou superfície para captura de imagens em uma determinada distância, onde não haja contato entre o imageador e a imagem (o que realiza a imagem e o que está sendo visualizado) (Divino, 2005). A captura da imagem é feita através de sensores que coletam a radiação refletida da Terra provinda do Sol até o sensor orbital. A sensibilização dos sensores gera diversos processos físicos para se obter, no final, uma imagem, que é uma representação do que o satélite estava "visualizando" em um instante de tempo da cobertura da Terra.

Desta forma, uma grande base de dados de SR é um conjunto de imagens e dados providos através de diversos satélites artificiais de Observação da Terra, utilizando técnicas de Sensoriamento Remoto, com diferentes resoluções espaciais e temporais, de diferentes tipos, complexidades e variedades, com tamanhos de arquivos extremamente massivos (Ma et al, 2015). Com isso, as imagens são capturadas enquanto os sensores ficarem ativos, sendo armazenados de forma exponencial. Além disso, com a evolução da tecnologia computacional e de sensores, a coleta dos dados através dos satélites tem se tornado cada vez mais um fluxo contínuo (Rathore et al, 2015), gerando dados em tempo real e transmitindo para os receptores quase que de forma imediata, por exemplo realizando streaming pela Internet. No final, esses dados capturados acabam gerando um conjunto de imagens de resolução temporal devido a frequência da passagem dos sensores em um mesmo ponto.

Quanto se trata de grandes bases de dados, hoje um conceito altamente discutido na literatura é o *Big Data*. E naturalmente a área de Observação da Terra, onde a produção de dados de medidas e imagens através dos satélites é maciça, adotou como uma nova metodologia de desenvolvimento, pois possibilitou uma nova maneira de tratar um conjunto massivo de informações (Guo et al., 2014; Tang and Liu, 2015). *Big Data* pode ser definido como uma ou mais grandes bases de dados de alta complexidade, heterogênea entre os tipos de dados nela contidos e em quantidade maciça, em que métodos e algoritmos tradicionais são ineficientes para tratá-los (Wand, 2014). Basicamente, a quantidade de dados é tão grande e complexa, com diferentes dados de diferentes fontes digitais como sensores, digitalizadores, modelos numéricos, smartphones, redes sociais entre outros (Yang et al, 2016), que sistemas tradicionais não conseguem processar tamanha quantidade de informações, seja pelo tempo de resposta para realizar uma operação ou como para simplesmente buscar o dado, por exemplo. Apesar do termo *Big* denotar um conjunto maciço de dados, nota-se que não existe uma definição de um tamanho específico para o que pode ser definido como *Big Data*, dado que o tamanho é apenas um dos parâmetros de análise. Além disto, o termo também é considerado muito amplo, devendo na verdade se focar na dificuldade de trabalhar com dados de diferentes dimensões (Agapiou, 2016), e não apenas no tamanho do conjunto de dados.

Para se definir um conjunto de dados como *Big Data*, foi estabelecido por Laney (2001) o que pode ser considerado a primeira definição de *Big Data* o conceito dos “três Vs”: Volume, Variedade e Velocidade. Volume se refere a quantidade dos dados; Variedade é a quantidade dos diferentes tipos de dados; e, por fim, Velocidade se trata da rapidez do processamento e transmissão dos dados. Há autores, porém, que determinam mais dois "Vs": Veracidade, que verifica a confiabilidade em relação a diversidade de qualidade, precisão e confiabilidade dos dados; e Valor, que se foca em pesquisas e decisões que melhoram a vida e o trabalho através destes dados (Mayer-Schönberger and Cukier, 2013), o quanto estes dados agregam valor para a sociedade. Por causa dessa heterogeneidade e alto volume, como também pela necessidade do rápido intercâmbio entre o imageador com as antenas terrestres, os dados de Sensoriamento Remoto entram na categoria dos "três Vs", e portanto são considerados como *Big Data* (Ma et al, 2015). Como o contexto de Veracidade e Valor não se aplica tanto aos dados de SR, tendo em vista que se espera que as fontes das imagens sejam os satélites de Observação da Terra e o valor é subjetivo, este trabalho considera apenas os três primeiros Vs.

De acordo com Rathore et al (2015), a análise de *Big Data* é um desafio devido a tarefa de localizar, identificar, entender e situar estes dados. Por causa da grande escala e dos diferentes tipos de dados (e até mesmo diferentes fontes de dados), se torna inviável também analisá-los através de modelos e algoritmos tradicionais (Agapiou, A., 2016), onde se espera que os dados sejam todos iguais ou, no mínimo, padronizados. Esta demanda se tornou tão grande que acabou gerando novas tecnologias focadas em armazenamento de grandes bases, como o banco de dados Hadoop e o paradigma de NoSQL (Yang, C. et al, 2016). Além disto, devido ao grande volume de dados, diversas áreas da ciência começaram a encontrar dificuldades em utilizar grandes quantidades de dados como a Geociência, Biologia, além do próprio Sensoriamento Remoto, entre outros (Wang, 2014).

Por tanto, diversas pesquisas foram realizadas para encontrar métodos que façam uso dessas novas estruturas de dados, buscando uma semântica entre as informações para se encontrar maneiras de avaliá-los de forma otimizada. Na literatura, foram propostas soluções para estes problemas, criando banco de dados específicos para *Big Data* e modelos de arquiteturas para serem utilizados em projetos estabelecidos para fazerem uso do grande volume de dados. Nos Capítulos seguintes serão apresentados com maiores detalhes alguma destas soluções, e também será discutido as características dos dados em relação a grandes bases do Sensoriamento Remoto.

2.2 Características para os dados das grandes bases de Sensoriamento Remoto (SR)

Uma das principais complexidades das grandes bases de SR não é somente o conteúdo da informação em si, mas também é relacionado a sua diversidade e alta dimensionalidade (Ma et al, 2015). Como grandes bases de SR podem ser consideradas como *Big Data* é necessário discutir sobre formas eficientes de armazenar, distribuir e tratar estes dados para que se possa fazer uso adequado das informações armazenadas.

Apesar de hoje em dia o armazenamento não ser mais um grande problema, tendo em vista que as máquinas conseguem ser modificados para receber mais memória física, a Observação da Terra produz uma quantidade imensa de dados com transmissões massivas na casa dos Terabytes todo dia, desde os satélites até os data centers onde estes dados serão armazenados

(Ma et al, 2015). Por exemplo, atualmente existem cerca de 200 satélites (Rathore et al., 2015) em órbita que capturam dados com resoluções multi-espaciais e multi-temporais. Com essa rede de satélites, utilizam-se técnicas de ciclo de menor caminho de revisita em um ponto para expandir a cobertura terrestre. Desta forma, gera-se um número bem maior de imagens da Terra, e consequentemente um maior número de dados. Analisando a partir de uma única base como exemplo, o NASA's Earth Observation System Data and Information System (EOSDIS) possui arquivos que ultrapassam 7.5 petabytes. Apenas durante o ano de 2012, mais de 4.5 milhões de gigabytes de dados foram distribuídos pelo EOSDIS. Desde 1972 os satélites da linha LANDSAT produzem imagens da superfície da Terra continuamente, com o Landsat 8 LDSM e Landsat 7 ETM+ adquirindo mais de 1200 imagens por dia (Agapiou, A., 2016). Com uma quantidade tão grande de dados, isto reflete em outra característica de grandes bases do Sensoriamento Remoto, que é a diversidade de tipos de dados.

Uma vez que os dados de Sensoriamento Remoto são utilizados para as mais diversas áreas da ciência da Terra como monitoramento ambiental, processos atmosféricos, hidrologia entre outros, torna o processo de análise dos dados mais complexa (Ma, Y. et al, 2015). Por exemplo, mais de 7000 tipos de conjunto de dados estão presentes nos arquivos da NASA, armazenadas em estruturas de arquivos dos mais diversos formatos como HDF, GeoTIFF, FAST entre outros. Devido aos diferentes formatos serem utilizados para ações específicas, com metatags e organizações de metadados distintos, isto aumenta a dificuldade em realizar processos de análises destes dados. Além disso, a distribuição destes dados também são importantes desafios. Como o hardware que realiza a coleta de imagens possui uma memória limitada, e a coleta de dados de SR ser algo constante, necessita de modelos que realizam persistência destes dados e, ao mesmo tempo, consiga uma memória escalável para não haver perda de dados. Desta forma, devido ao grande volume, transferir estes dados (seja do imageador até um ponto de recepção dos dados, como na rede interna ou Internet) também necessitam de estratégias importantes, com algoritmos de compressão eficientes como uma maneira de pré-processamento para que possam diminuir o tamanho dos dados antes de transferi-los. Por exemplo, uma técnica proposta por Li et al. (2015) organiza um modelo de transmissão de rede com técnicas de compressão de dados para tramitação de dados geoespacial em um ambiente de cyberinfraestrutura.

Apesar da quantidade de dados de SR aumentar a cada dia, todas estas imagens são de alto valor para a Observação da Terra, não podendo serem simplesmente descartados. Elas apresentam anos de transformações do planeta e possibilitam estudos através de séries temporais por aplicações do Sensoriamento Remoto, como por exemplo para verificar deformações na superfície e classificação da cobertura de terra (Lin, F.C., 2013), além de aplicações de monitoramento de áreas florestais, falecimentos, emergências. Por tanto, modelos de persistência e acesso rápido aos dados através de buscas eficientes para bases de *Big Data* são essenciais para tornar os dados de SR mais acessíveis para poderem ser utilizadas em políticas públicas e desenvolvimento tecnológico. Para isto, *Cloud* e NoSQL são dois paradigmas na área de Tecnologia da Informação que se tornaram soluções essenciais para o conceito de *Big Data*, e que possibilitou a acessibilidade dos dados e apresentou um meio de processamento mais eficiente aos dados.

Cloud pode ser definido como uma virtualização de máquinas descentralizadas, onde diferentes computadores são distribuídas por nós e não precisam necessariamente estarem em um mesmo local. Com essa distribuição, a comunicação é realizada através da Internet ou por uma rede interna própria. Uma das vantagens deste esquema é que possibilita que o sistema

seja altamente escalável, podendo aumentar o número de hardware quando for necessário (seja ela para memória, processamento ou de I/O) e realizar a comunicação destes componentes paralelos através de rede ou pela internet. Além disto, permite que os dados estejam distribuídos em diferentes nós, o que possibilita que os dados não precisem estar em um mesmo hardware, mas divididos em diferentes nós. Esta virtualização permite que a supercomputação para processamento de massas de dados se torne mais acessível (Ma, Y. et al, 2015), aumentando o número de nós interligados à rede quando necessário em vez de simplesmente trocar de máquina ou aumentar a quantidade de memória de uma única máquina, por exemplo.

Para fazer fruto da *Cloud*, o NoSQL aparece como uma alternativa para persistência que utiliza do conceito de distribuição paralela. Definido como "Not Only SQL" (Não Só SQL), é uma extensão da linguagem SQL para banco de dados focados no controle de grandes estruturas e dados não relacionados (Ma, Y. et al, 2015). O NoSQL se foca menos na estrutura da informação e mais no armazenamento dos dados, diferente do relacional em que os dados são armazenados em tabelas com estruturas de dados (esquemas) fixas. Desta forma os dados podem ser inseridos no banco de dados sob demanda à medida que o sistema receber este dado, não importando o esquema das tabelas. Por isso, é dito que banco de dados NoSQL são considerados de esquema livre. Outra característica para Banco de Dados NoSQL é a sua abordagem de armazenamento por chave-valor, o que traz alta velocidade na busca e alta escalabilidade.

Por causa da praticidade da *Cloud* e velocidade no armazenamento e busca dos banco de dados NoSQL, os dois conceitos se tornaram meios importantes para o tratamento de dados de bases *Big Data*. Através da *Cloud*, gera-se a virtualização para o processo de tratamento das imagens de SR, além do armazenamento em diferentes máquinas para não sobrecarregar a memória de um hardware. Já o NoSQL possibilitou um meio de controle destes conjuntos de dados nos diferentes nós, tratando os dados conforme o esquema dos tipos destes dados forem sendo gerados, o que é importante dado a natureza heterogênea dos dados de SR. Um importante exemplo de aplicação para o armazenamento e acesso aos dados é o portal GEOSS, um conjunto de sistemas coordenados e independentes de observação, informação e processamento da Terra que interagem e fornecem acesso a informações diversas para uma ampla gama de usuários do setor público e privado. O GEOSS liga esses sistemas para fortalecer o monitoramento do estado da Terra, facilita o compartilhamento de dados ambientais e informações recolhidas a partir da grande variedade de sistemas de observação contribuídos por países e organizações pertencentes ao GEO (Group on Earth Observations). Os fornecedores dos dados são instituições de países pertencentes ao GEO de toda a comunidade internacional, inclusive o Instituto Nacional de Pesquisas Espaciais (INPE).

Apesar dos esforços de adaptar os sistemas da Observação da Terra, estes não conseguem nativamente tratar as grandes bases de dados por requerer uma Engenharia de Software nova para o sistema e a re-adaptação do sistema, podendo levar muito tempo. Portanto, para que isso seja possível usar estas arquiteturas, são aplicadas nestes softwares módulos de visualizações que possui arquiteturas específicas para tratar e extrair informação destas grandes bases. Estes módulos são então usados em conjunto com os sistemas. No Capítulo a seguir será descrito com maiores detalhes algumas arquiteturas propostas pela literatura para os módulos de visualizações.

2.3 Estratégias existentes para os módulos de visualizações

Diversas arquiteturas foram propostas na literatura como formas de melhor aproveitar os recursos que dados considerados como *Big Data* podem oferecer, desde a tomada de decisões quanto a manipulação de grandes números de informações. Contudo, conforme a abordagem foi melhor estudada, propostas mais eficientes visando uma arquitetura mais genérica e que faça uso de banco de dados distribuídos foram considerados. A seguir serão expostos três trabalhos, por ordem de ano, que sugerem arquiteturas como formas de aproveitar as grandes bases de dados de acordo com cada linha de pesquisa sugerida. A primeira é um trabalho de dos Santos et al (2013) de uma arquitetura baseada em nuvem para Tempo e Clima, proposto para um trabalho em conjunto com o CPTEC. A segunda, no subcapítulo 2.3.2 é focada em análise dos dados para aplicações do SR sugerida por Rathore et al (2015), enquanto que no 2.3.3 é uma arquitetura de multi-usuário que visa a eficiência de uma análise próxima dos dados.

2.3.1 Uma arquitetura privada baseada em Cloud para uma Observatório Virtual Brasileiro de Meteorologia e Clima

dos Santos et al (2013) descreve em seu trabalho uma arquitetura voltada para um Observatório Virtual privado baseado em Cloud para dados meteorológicos e climáticos. Este trabalho foi uma parceria entre o INPE e o CPTEC que iniciaria um projeto de centralização dos dados meteorológicos e climáticos, e disponibilizar na internet. Desta forma, o usuário poderia visualizar os dados climáticos e sua relação temporal, como por exemplo quantas vezes ocorreu uma tempestade em uma determinada época ou qual foi o maior período de seca.

Observatórios Virtuais (OV) são arquiteturas capazes de organizar, manter e explorar grandes informações, distribuídas e de tipos de dados dinâmicos. Através desta arquitetura é possível uma maneira de catalogar dados, processá-los, visualizá-los e relacioná-los através de ferramentas tanto web quanto desktop. Com isso, a arquitetura proposta possui um conjunto de ferramentas de software de acesso a dados que permitirão aos usuários de diferentes níveis e conhecimentos descobrir dados, fazer análises e visualizações básicas, usando protocolos de acesso a dados uniformes via serviços web, disponível em uma Cloud privada. A Figura 3 mostra de forma geral o modelo da arquitetura para o OV, que é agrupada em três principais categorias: servidor banco de dados e aplicações baseada em Cloud privada, servidor de dados externos e clientes externos.

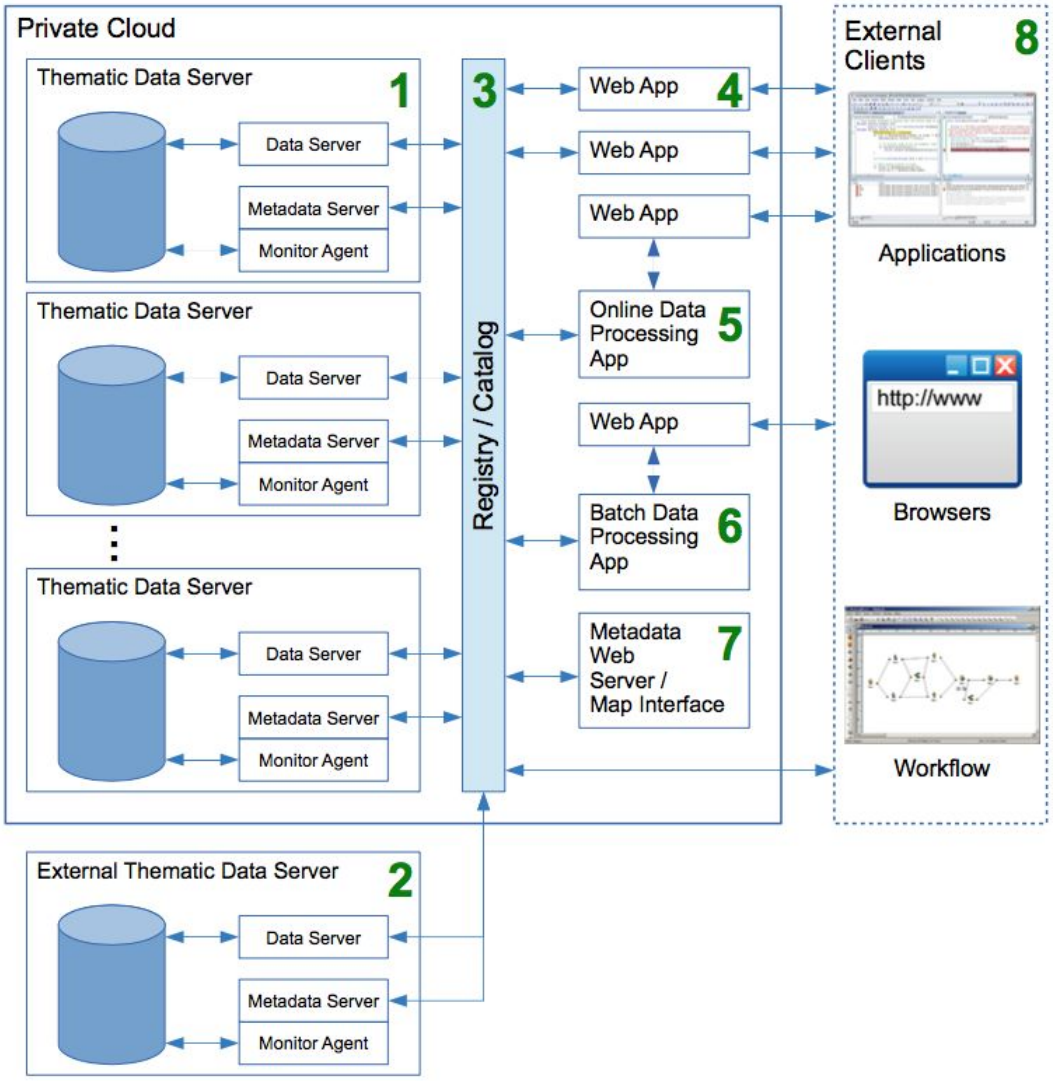

Figura 3 Modelo da arquitetura para o Observatório Virtual

A Cloud privada é composta por um servidor de dados temáticos, um registrador/catálogo, a aplicação web, a aplicação de processamento de dados online, uma aplicação de processamento de dados em lote e um servidor de metadados web/interface map.

- Servidor de dados temáticos: possui dados de uma coleção de dados específica ou uma generalização a partir de um tema específico (como temperatura, descarga elétrica na atmosfera, velocidade e direção do vento, entre outros), em que estes dados podem ser espacialmente e temporalmente limitadas. O servidor de dados temáticos também possui um agente de monitoramento, que regulariza as queries do banco de dados contido no servidor de dados temáticos para extrair metadados, e um servidor de metadados, que armazenará as informações extraídas (em contexto de OV, um metadado são informações sobre clima e meteorologia armazenados no banco de dados, em particular, informações sobre a cobertura de dados (a extensão espacial ou temporal desses dados).
- Registrador ou Catálogo: repositório central da informação sobre dados, metadados e ferramentas para a OV que podem ser buscadas a partir de coordenadas geográficas, intervalos de tempo, tipos de dados, palavras chaves, entre outros, que retorna um conjunto de recursos (normalmente serviços web) que podem ser usadas para

conseguir o dado. A interação com o registrador pode ser feita via interface web ou serviço web.

- Aplicação web: interfaces de registro e banco de dados associados ao OV, implementadas como um serviço web que executa queries no banco de dados ou registradores e retorna o resultado para a aplicação cliente. A aplicação web também pode compor resultado do banco de dados ou de outras aplicações, assim como distribuir e agregar queries, executando algoritmos específicos. Desta forma, a aplicação web consegue executar queries mais complexas e algoritmos mais complexos.
- Aplicações de processamento de dados online: aplicações para consultar e processar um ou mais servidores de dados e/ou registradores para gerar resultados quase em tempo real. Essas aplicações irão interagir com usuários e aplicativos externos através de aplicativos web específicos que funcionam como portais para os aplicativos de processamento de dados, ou seja, interfaces que podem chamar as aplicações, passar parâmetros e retornar os resultados ao usuário final.
- Aplicação de processamento de dados em lote: aplicações que fará o acesso aos servidores de dados e registradores, processa estes dados e retorna ao usuário a partir de aplicações web específicas que controlam a execução das aplicações de processamento de dados. No entanto, a execução da aplicação é realizada em lote, ou seja é possível rodar algoritmos mais intensos mas sem a garantia de que os resultados serão entregues em tempo real. Técnicas de processamento em lote são aplicadas, como filas de prioridade e processos de execução e monitoramento, por exemplo.
- Servidor de metadados web/Interface Map: aplicação web que permite a descoberta visual dos dados disponíveis através de um serviço web, que poderá interagir com outras aplicações e listar as fontes de dados correspondentes a uma restrição específica (tipo de dado, tempo e limite espacial, qualidade do dado, etc.), e por uma interface gráfica, que apresenta os resultados para fontes sobrepostas em um mapa semelhante ao SciScope. Desta forma, usuários poderão localizar regiões rapidamente em tempo e espaço que contenham os dados de interesse.

O segundo componente principal da arquitetura é o servidor de dados externos temáticos, que basicamente é um banco de dados de colaboradores fora da Cloud privada, disponível em outra localização. Estes servidores externos precisam implementar o Agente Monitor e o Servidor de metadados para poderem ser reconhecidos pelo Observatório Virtual, porém o banco de dados não necessita de qualquer alteração. Isto torna o processo mais orgânico, já que pesquisadores não precisam alterar a arquitetura do banco de dados, apenas criar um link entre o banco de dados e o OV. O terceiro e último componente principal são os clientes externos, que serão as aplicações e serviços que farão uso dos metadados do OV. Espera-se que diversos tipos de clientes façam requisições do tipo serviços web ao OV, como via Browsers, aplicativos, workflow, etc.

Conforme descrito na Figura 3, a funcionalidade básica do OV (alguns dos seus servidores de dados, o registrador, alguns aplicativos) será implantada em uma nuvem privada, para ser operado por uma única empresa ou instituição. dos Santos (2013) descreve diversas vantagens em operar numa nuvem privada, tais como: as aplicações web e registros serão hospedados em um mesmo ambiente físico conectada por uma rede interna, o que provê alto desempenho; servidores temáticos possuem recursos em comum, desta forma implantar dados em novos servidores virtuais seria mais fácil; e o monitoramento do desempenho dos servidores de

dados virtuais e aplicações web podem fornecer informações sobre o uso, o que pode ajudar a mudanças nos recursos alocados no servidor.

2.3.2 Arquitetura de análise de Big Data em tempo real para aplicações do Sensoriamento Remoto

Rathore et al (2015) propôs uma arquitetura de análise de Big Data em tempo real para aplicações do Sensoriamento Remoto que processa e analisa em tempo real e offline dados do SR para tomada de decisão. Esta arquitetura foi implementada através da linguagem Java usando Beam-5.0 e o Hadoop usando MapReduce. De acordo com Rathore, este modelo de arquitetura realiza uma análise de uma maneira eficiente ao separar em unidades as etapas de aquisição, processamento, e análise e decisão. A Figura 1 mostra de maneira ilustrativa os componentes desta arquitetura, enquanto que a Figura 2 mostra o fluxograma de passos desta arquitetura. De forma geral, a comunicação da infraestrutura desta arquitetura se dá inicialmente pela unidade de aquisição (Remote sensing big data acquisition unit, RSDU), que irá obter através de um número *n* de satélites as imagens de observação da Terra através dos sensores, com as cenas sendo gravadas através da radiação refletida até os satélites e então transmitidas até as bases de estação na Terra, sendo pré-processadas. Com isso, estes dados são filtrados e efetivamente processados na unidade de processamento dos dados (DPU), e analisados na unidade de análise e decisão dos dados (DADU).

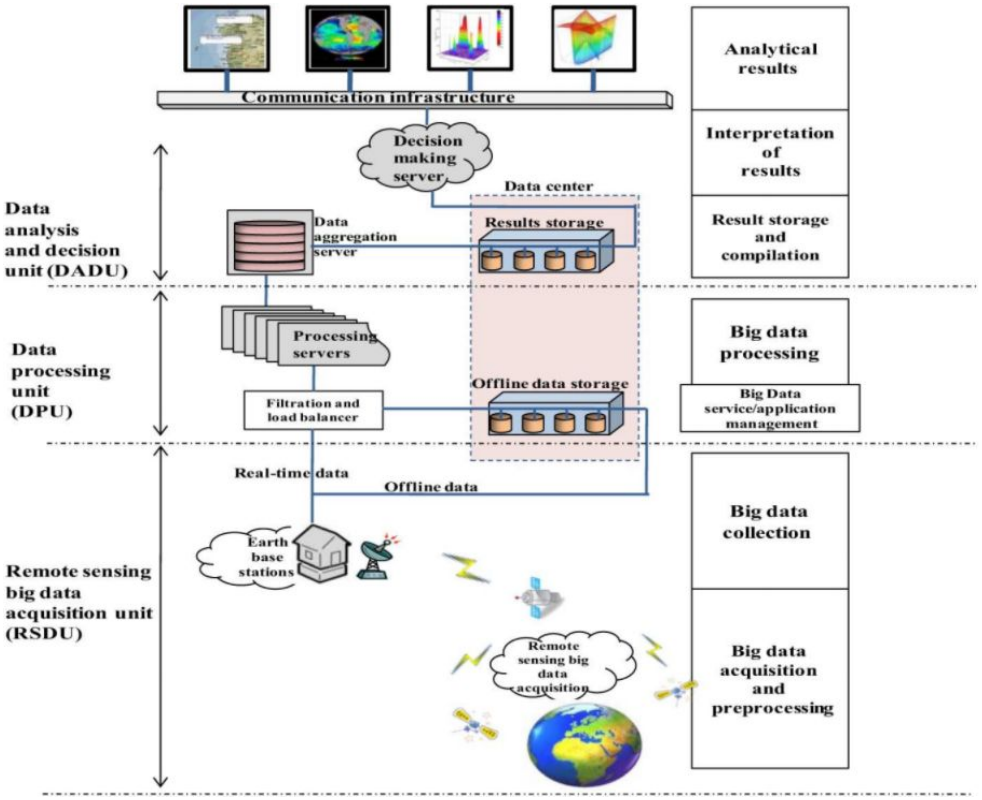

Figura 1 Arquitetura de Big Data para Sensoriamento Remoto

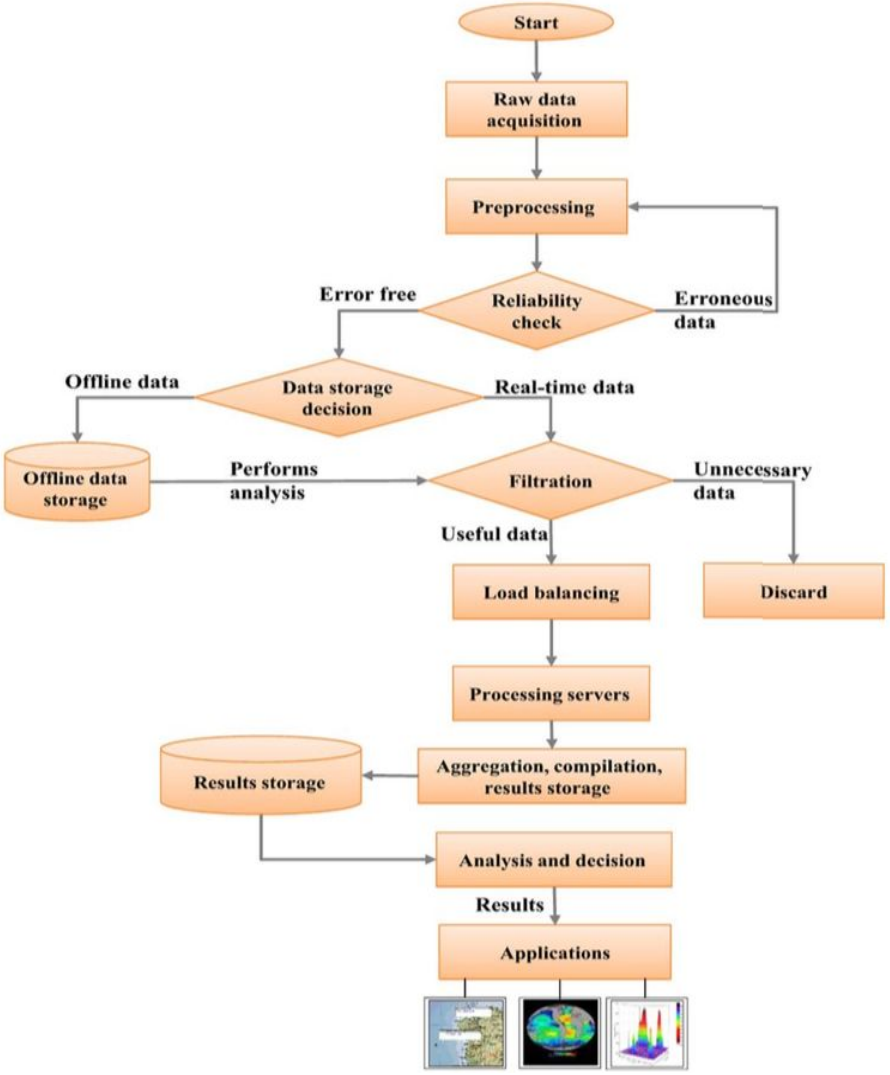

Figura 2 Fluxograma da Arquitetura de análise de dados do Sensoriamento Remoto

Na unidade de aquisição de *Big Data* de Sensoriamento Remoto (RSDU), a aquisição das imagens é realizada de forma paralela reunindo os dados de diversos satélites em órbita da Terra. Com a massa de dados recebida, estes são filtrados e pré-processados. Devido a natureza maciça de dados, em que tecnologias atuais não conseguem processar essas imagens de maneira eficiente por falta de energia necessária, o pré-processamento também é realizado de forma paralela. Além disso, apesar das imagens podem estar distorcidas, devido a dispersão e absorção de gases atmosféricos e partículas de poeira que podem ter interferido no registro da imagem, esta etapa (RSDU) assume que o satélite consegue corrigir esses erros. Após tornar os dados recebido em um formato imagem, utilizando algoritmos de Doppler ou SPECAN, o pré-processamento efetivo utiliza-se algumas técnicas como integração de dados, limpeza de dados e a eliminação de redundância. Isto é importante pela fonte dos dados serem de diversos satélites em órbita e geralmente estes dados são integrados, o que pode ocorrer desde a repetição de locais ou até imagens com problemas. No final, isso acaba dando mais precisão para a análise destes dados. Então, estas imagens são transmitidas para uma estação terrestre via canal Downlink que trata estas imagens e corrigir possíveis distorções causados devido ao movimento ou curvatura da Terra, iluminação, variação das características do sensor, entre outros. Por fim, as imagens são transmitidas para a Estação Base da Terra para processamento posterior. Na transmissão, as imagens são enviadas de duas formas paralelas:

1) processamento de grandes dados em tempo real, em que as imagens são transmitidas diretamente para os servidores de filtração e balanceamento de carga, pois armazená-los primeiro iria degradar o desempenho de processamento em tempo real; e 2) processamento de grandes dados offline, onde a estação base terrestre transmite para um datacenter de armazenamento em que estas imagens serão utilizadas para futuras análises.

A segunda parte da arquitetura é composta por uma Unidade de Processamento de Dados (Data Processing Unit - DPU), que contém a etapa de filtrar e balancear a carga. Assim sendo, ela tem duas responsabilidades principais: o filtro identifica o dado útil para uma determinada análise, bloqueando e descartando dados em que não é possível extrair informações relevantes para a análise útil do problema, resultando em um melhor desempenho no sistema; e o balanceamento de carga do servidor fornece facilidade na divisão dos dados filtrados em partes, distribuindo o processamento para vários servidores, não sobrecarregando-os. Desta forma, o algoritmo de filtração e balanceamento de carga varia de análise para a análise. Cada servidor fará cálculos estatísticos, de medição e realiza outras tarefas matemáticas ou lógicas para gerar resultados intermediários para cada segmento dos dados. Os resultados gerados são enviados para o servidor de agregação para compilação, organização e armazenamento para posterior processamento.

A terceira e última parte da arquitetura é a Unidade de Análise de Dados e Decisões (Data Analysis and Decision Unit - DADU). Esta parte contém três grandes componentes formados por servidores: servidor de agregação e compilação, servidor(es) de armazenamento de resultados e servidor de tomada de decisão. O fluxo se inicia a partir da DPU no servidor de processamento, que envia os resultados parciais para o servidor de agregação e compilação, que irá agregá-los a partir de algoritmos que compilam, organizam, armazenam e transmitem os resultados. Estes algoritmos também irão variar dependendo da necessidade da análise. O servidor de agregação e compilação armazenará os resultados compilados e organizados para que qualquer servidor possa usá-los durante o processamento a qualquer momento, e uma cópia é enviada para o servidor de tomada de decisão, que usa algoritmos de decisão para questionar diferentes questões de um determinado resultado para, em seguida, realizar decisões (por exemplo, encontrar pontos de terra, mar, fogo, tempestade, entre outros). Por fim, exibe o resultado ou transmite a decisão, fazendo com que qualquer aplicação (seja software comercial ou até redes sociais, por exemplo) possa receber este dado em seu processamento paralelamente.

2.3.3 SciServer Compute

Por fim, a última arquitetura proposta é a SciServer Compute (Medvedev et al, 2016), uma infraestrutura modular e escalável de dados para o armazenamento, acesso, consulta e processamento de bases científicas *Big Data* na escala dos Petabyte. Possui diversas infraestruturas altamente integradas que trabalham em conjunto para formar um sistema completo, possibilitando a análise de grandes bases de dados científicos sem realizar o download/upload desses dados. Além disto, pode-se inserir algoritmos de análise através do notebook online Jupyter, que funciona em contêineres ao lado do servidor local e que estão conectados as grandes bases de dados relacionais armazenados, tornando o processo de análise mais ágil. O objetivo principal do SciServer Compute é fornecer um espaço de trabalho computacional que executa o código do usuário para a análise dos dados de forma eficiente.

De acordo com Medvedev et al (2016), diversas metas foram projetadas na construção da arquitetura do SciServer. 1) acessar e processar grandes bases de dados que seriam impossíveis de fazer o download destes dados. 2) ter armazenamento personalizado para os usuários do sistema para que pudessem armazenar tanto o resultado das consultas, quanto o resultado de processos adicionais dos arquivos armazenados; também ser possível realizar uma consulta cruzada, e processar o próprio conjunto de dados em repositórios próprios. 3) ter um ambiente de compartilhamento colaborativo de dados por grupos de usuários. E, por fim, 4) o principal objetivo foi prover uma arquitetura de sistema com capacidade de computação escalável para o processo e análise dos dados.

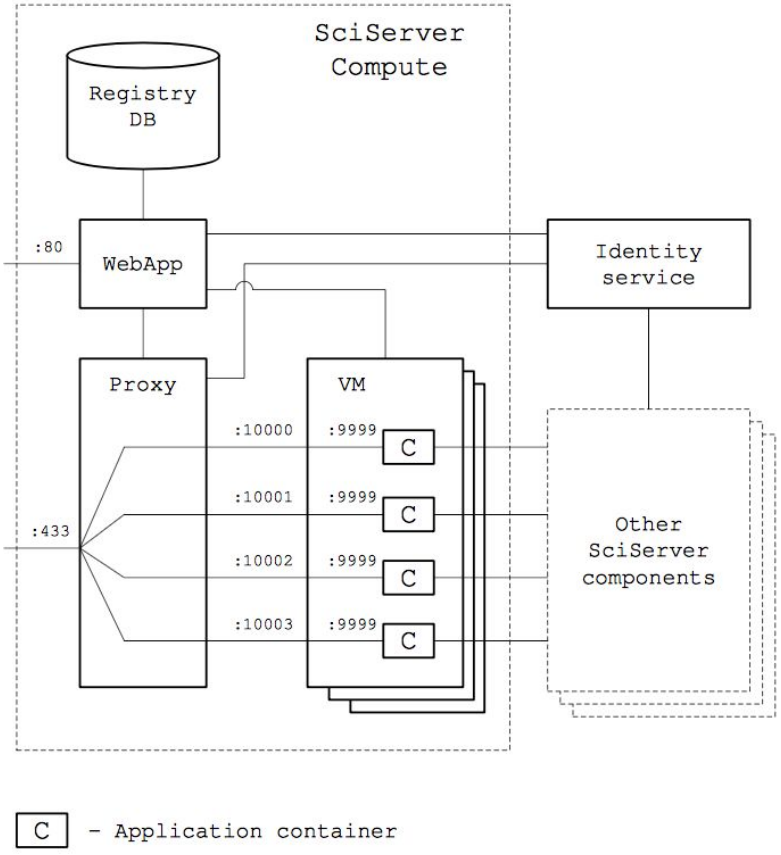

Figura 4 Arquitetura geral do SciServer Compute

Apesar do projeto ser focado para integrar grandes bases de dados existentes nos campos da astronomia, cosmologia, turbulência, genômica, oceanografia e ciências materiais, é possível adaptá-la para outros campos de pesquisa que necessite alto processamento e desempenho como a Observação da Terra, para análise dos dados de Sensoriamento Remoto, através da inserção de módulos e bases de dados que lidem com estes tipos de dados. Isto ocorre porque a arquitetura trabalha na forma de módulos, que executam scripts no notebook Jupyter em uma máquina virtual (MV) que possui um número limitado de containers Docker (API de alto nível para gerenciamento de contêineres em sistemas Linux). Os nós, contêineres de aplicativos e os volumes de dados são orquestrados do lado do servidor, que funcionam como parte do WebApp e que faz uso do banco de dados de registro global (Registry DB). Esta abordagem oferece escalabilidade e flexibilidade suficientes para o sistema.

A arquitetura do SciServer Compute é composta pelos seguintes componentes: 1) Registry DB, que contém as grandes bases de dados *Big Data* para serem analisados; 2) WebApp, que fará os processos através de requisições web; 3) Proxy, que fará a autenticação dos usuários ao servidor; 4) VM, que contém as aplicações rodando o notebook Jupyter em containers de cada usuário paralelamente; 5) Identity service, que fará a identificação no servidor; e 6) Other SciServer components, que são outros componentes inseridos de forma escalável em que as aplicações poderão realizar os processos. O objetivo do SciServer Compute é fornecer aos usuários seus próprios espaços de trabalho computacionais para executar seu código para análise de dados, e isto é provido através da API Docker para o gerenciamento de containers em sistemas Linux. Cada container é hospedado em uma Máquina Virtual que é acessado por um determinado usuário, e cada Máquina Virtual possui um notebook Jupyter para o usuário entrar com sua própria aplicação a ser rodada via web. Esta aplicação terá acesso local ao Registro de banco de dados global para realizar um determinado processo. Como tudo isto é realizado no servidor via web, torna o processo mais eficiente pois não é necessário enviar ou baixar dados.

3. Crítica e Conclusão

Neste trabalho foram mostrados três exemplos de arquiteturas para o uso e tratamento das grandes bases de dados do Sensoriamento Remoto para diferentes propostas de análise de dados. Os três trabalhos propuseram arquiteturas diferentes para o gerenciamento e processamento de dados, desde compor o ambiente de captura e tratamento das imagens orbitais até no foco em processamento de alto desempenho através do uso de nuvem privada. Enquanto um é mais específico para os dados da Observação da Terra, como é o caso de Rathore et al (2015), outras duas são aplicadas para dados de análises mais generalizados que, devido a forma como foram implementadas, podem ser aplicadas aos dados de SR, como é o caso de dos Santos et al (2013) e Medvedev et al (2016).

Um ponto importante sobre esta pesquisa foi mostrar trabalhos de diferentes épocas e que, mesmo com uma pouca diferença de tempo entre elas, mostra como a área de *Big Data* tem sido extensivamente estudado durante os anos e tendo uma rápida evolução em questão de conceitos e tecnologias. Naturalmente, o trabalho de Medvedev et al é o mais avançado em termos de tecnologia e integração com a nuvem, apresentando uma arquitetura de desenvolvimento totalmente integrada a nuvem, porém analisando em paralelo as propostas de dos Santos e Rathore, conclui-se que a área de *Big Data* pode estar caminhando para pesquisas com foco em se ter mais nuvens privadas para processamento em servidor local, onde a internet não seja a única fonte para comunicação entre os banco de dados.

Com isso, também conclui-se que o principal problema para o as grandes bases de dados é o deslocamento destes dados, devido a quantidade exorbitante que impede o upload/download pela rede. Devido a isso, foi-se necessário criar componentes paralelos que pudessem ser inseridos de forma escalável e sob demanda do problema, principalmente em questão de análise e tratamento específico de uma imagem orbital, para criar sistemas capazes de utilizar recursos de maneira eficiente. E para isso, este trabalho indica também que um caminho é a criação de nuvens privadas para diversas fontes de dados (seja ela local e pela internet) em um ambiente que o usuário do sistema acrescente componentes para o processamento dos dados conforme o problema necessitar (ou seja, programar no próprio ambiente e inserir

componentes), tudo isso integrado a um conjunto de servidores locais para gerar alto desempenho.

4. Referências

Figueiredo, Divino. "Conceitos básicos de sensoriamento remoto." *Companhia Nacional de Abastecimento-CONAB. Brasília-DF* (2005).

Guo, H., Wang, L., Chen, F., Liang, D., 2014. Scientific big data and digital earth. Chin. Sci. Bull. 59 (12), 1047e1054.

Laney, D., February 2001. 3D Data Management: Controlling Data Volume, Velocity and Variety. Gartner. Available at: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

Lin, Feng-Cheng, et al. "The framework of cloud computing platform for massive remote sensing images." *Advanced Information Networking and Applications (AINA), 2013 IEEE 27th International Conference on*. IEEE, 2013.

Ma, Y., Wu, H., Wang, L., Huang, B., Ranjan, R., Zomaya, A., & Jie, W. (2015). Remote sensing big data computing: Challenges and opportunities. *Future Generation Computer Systems*, *51*, 47-60.

Mayer-Schönberger, V., and K. Cukier. 2013. Big Data: A Revolution That Will Transform How We Live, Work, and Think. Houghton Mifflin Harcourt. (Chapter 1).

Medvedev, D., Lemson, G., & Rippin, M. 2016. SciServer Compute: Bringing analysis close to the data. In *Proceedings of the 28th International Conference on Scientific and Statistical Database Management*, eds. P. Baumann, I. Manolescu-Goujot, L. Trani, G. G. Barnafoldi, L. Dobos, & E. Banyai. New York:ACM. 27.

M. M. U. Rathore, A. Paul, A. Ahmad, B. W. Chen, B. Huang and W. Ji, "Real-Time Big Data Analytical Architecture for Remote Sensing Application," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 10, pp. 4610-4621, Oct. 2015.

Nativi, Stefano, et al. "Big data challenges in building the global earth observation system of systems." *Environmental Modelling & Software* 68 (2015): 1-26.

Tang, Z., Liu, T., 2015. Evaluating internet-based public participation GIS (PPGIS) and volunteered geographic information (VGI) in environmental planning and management. J. Environ. Plann. Man. http://dx.doi.org/10.1080/ 09640568.2015.1054477.

Wang, Lizhe, et al. "Estimating the statistical characteristics of remote sensing big data in the wavelet transform domain." *IEEE Transactions on Emerging Topics in Computing* 2.3 (2014): 324-337.

Yang, C., Q. Huang, Z. Gui, Z. Li, C. Xu, Y. Jiang, and J. Li. 2013. “Cloud Computing Research for Geosciences.” In Spatial Cloud Computing: A Practical Approach, edited by C. Yang, Q. Huang, Z. Li, C. Xu, and K. Liu, 295–310. Boca Raton, FL: CRC Press/Taylor & Francis.