Capítulo 15

Métodos, procedimentos e técnicas utilizadas na construção de AgroTIC

Leandro Carrijo Cintra
Alan Massaru Nakai
Jorge Luiz Correa

1 Introdução

Com o afloramento do método científico no século 16 ficou evidente que a humanidade teria enormes benefícios se tratasse os problemas agrícolas com um procedimento formal e replicável. Iniciou-se então um ciclo que perdura até o presente, no qual muito se avançou no conhecimento sobre cultivo de plantas e criação de animais, além do domínio dos mais variados microrganismos em diversos processos relacionados à agricultura e à alimentação.

A ciência desempenhou um papel fundamental e extremamente relevante para que a humanidade alcançasse o estágio de desenvolvimento em que se encontra nas ciências agrícolas; e atualmente há projetos de pesquisa que permeiam absolutamente todos os pontos de interesse no que diz respeito às questões agropecuárias. Apenas para citar alguns exemplos, sem a mínima intenção de exaurir toda a lista de áreas de atuação, pode-se mencionar projetos no estudo de solos; fisiologia de plantas e animais; nutrição animal e vegetal; genética e melhoramento de microrganismos, animal e vegetal; seleção genômica; rastreabilidade; monitoramento por satélite e veículo aéreo não tripulado (VANTs) aplicados à agricultura; modelagem agroambiental, impactos da agricultura no clima e impactos da mudança climática na agricultura.

O que todos estes projetos têm em comum é o fato de se basearem sempre no método científico para estabelecer os seus resultados. Método este que compreende um conjunto de passos que ajudam a formular, corrigir e evoluir teorias sobre observações e problemas de interesse.

Atualmente, o método científico tem sofrido uma revolução em função do uso massivo da tecnologia da informação (TI) nas mais diversas etapas. A etapa de experimentação, por exemplo, tem usado intensivamente a TI para a geração de dados e informações acerca de eventos sob estudo em várias circunstâncias. Com a criação de equipamentos eletrônicos e softwares especializados em capturar dados, têm-se gerado vultuosos volumes de dados para posterior análise e obtenção de resultados em várias áreas do conhecimento. Em diversas situações estes repositórios são referenciados como *big data*. A quantidade de dados gerados é tamanha que torna impossível a análise dos mesmos sem o uso de processamento computacional intenso, e novamente componentes da tecnologia da informação aparecem associados a outra etapa do processo científico, tornando a TI essencial para o avanço de uma ampla gama de áreas do conhecimento humano.

A relação entre o processo científico e a tecnologia da informação tem se tornado tão estreita que alguns especialistas advogam que se está vivenciando uma nova era nas ciências, na qual a exploração de dado é colocada como um novo paradigma na ciência moderna (HEY et al., 2011). Assim, há aproximadamente mil anos, a ciência era puramente empírica, baseada na descrição de fenômenos naturais; há poucos séculos, tornou-se teórica, baseada em modelos e generalizações; nas últimas décadas ficou explícito o paradigma computacional da ciência, baseado em simulações de fenômenos complexos; e atualmente, fala-se no paradigma baseado na exploração de dados, comumente referenciado como e-Science.

Independentemente das discussões sobre a atual configuração dos paradigmas científicos, é evidente que na atualidade as ciências não poderiam abdicar do uso da computação para alcançar as suas metas. Isto tem ocorrido também com uma ampla gama de estudos relativos à agricultura.

É neste ponto que se torna indispensável uma infraestrutura computacional adequada para fazer frente aos diversos grandes desafios que se apresentam atualmente no âmbito das pesquisas que permeiam o campo da agricultura. Para evidenciar esta demanda, pode-se citar dois casos, dentre vários outros, em atividades de projetos na Embrapa Informática Agropecuária. No primeiro, um projeto para sequenciamento do *Bos Indicus* (popularmente conhecido como Nelore), uma espécie com alto interesse comercial, gerando um volume de dados inicial da ordem de 1,5TB (um e meio terabytes). Durante a execução do projeto atingiu-se um patamar de 15TB de dados intermediários, tendo utilizado na fase crucial de montagem do genoma 500GB de memória RAM e 96 threads de processamento por um período de 3 dias consecutivos. O segundo caso está relacionado com o processamento e análise de séries temporais para avaliação de modelos em mudanças climáticas. Um projeto nesta linha de ação estima a necessidade de armazenamento temporário de 100TB para a execução de suas atividades.

Estes exemplos evidenciam a necessidade efetiva do uso de soluções computacionais de alto-desempenho nas atividades de pesquisas agrícolas atualmente. Sendo assim, segue-se uma discussão sobre as principais infraestruturas computacionais utilizadas para suporte aos diversos projetos de pesquisa realizados na Embrapa Informática Agropecuária.

2 Arquiteturas computacionais aplicáveis a problemas científicos

A principal característica exigida dos sistemas computacionais que apoiam os projetos científicos na Embrapa Informática Agropecuária é a escalabilidade. Esta característica diz respeito à capacidade que os sistemas têm de crescerem em relação a alguma de suas propriedades, tais como, capacidade de processamento, tamanho de memória RAM e capacidade de armazenamento. Quando se trata de escalabilidade, os sistemas podem encaixar-se em duas categorias: escaláveis verticalmente ou horizontalmente. Na primeira categoria estão as máquinas que individualmente podem atingir milhares de processadores, possuir dezenas de terabytes[1] de memória RAM e

[1] Terabytes: unidade de medida de volume de dados em um sistema computacional. Segundo o sistema internacional de medidas equivale a 10^{12} bytes (caracteres).

sistemas de armazenamento especializados que podem atingir petabytes[2] de capacidade. Na segunda categoria encontram-se os *clusters* e grids[3], que permitem a ampliação dos seus recursos com a adição de novos "nós", ou seja, novas máquinas. O custo por processador ou por unidade de memória é mais expressivo nas máquinas escaláveis verticalmente, porém existem problemas que somente são bem equacionados em arquiteturas deste tipo. Na Embrapa Informática Agropecuária tem-se uma infraestrutura baseada nas duas arquiteturas.

Outro ponto vital relacionado ao tema infraestrutura computacional diz respeito à forma como os recursos são alocados aos usuários. Na atualidade, a computação em nuvem apresenta uma abordagem bastante prática para esta questão, o que tem motivado as organizações a investirem esforços e recursos financeiros nesta área. Um dos projetos da Embrapa Informática Agropecuária tem investigado o uso desta tecnologia aplicada aos problemas relacionados à infraestrutura que surgem comumente nas pesquisas agropecuárias.

2.1 Arquitetura altamente escalável

O Laboratório Multiusuário de Bioinformática da Embrapa (LMB), que possui sua infraestrutura computacional hospedada na Embrapa Informática Agropecuária, é um exemplo de arquitetura escalável. Sua missão é propiciar uma infraestrutura computacional moderna e atualizada que faça frente aos desafios relacionados à bioinformática na Embrapa. A bioinformática, por sua vez, trabalha com o armazenamento, a organização e o processamento de dados biológicos, os quais, atualmente advém de várias fontes. A principal demanda dos problemas desta área é por máquinas com grande capacidade de processamento e que tenham memória RAM considerável. Em virtude destas demandas, o laboratório utiliza uma infraestrutura escalável nas duas frentes, ou seja, tanto vertical quanto horizontalmente. A Figura 1 ilustra o ambiente computacional destinado ao processamento de dados biológicos na Embrapa.

O ambiente computacional é constituído por dois storages para armazenamento. O primeiro é um *Storage Area Network* (SAN) com capacidade para armazenar 100TB de dados. O segundo é um *Network-attached Storage* (NAS) com capacidade para armazenar 220TB, sendo expansível a até 1,3PB (petabytes). Este é um exemplo de máquina com grande escalabilidade vertical.

Para atender às demandas por processamento com boa capacidade de memória, tem-se um grid computacional com a maioria dos nós de processamento possuindo quantidade mediana de memória e alguns poucos nós especiais, com grande volume. Assim, a capacidade de processamento pode ser expandida horizontalmente com a adição de novos nós simples. Já a capacidade de memória disponível para a solução de um mesmo problema pode ser expandida com a adição de nós especiais com alta capacidade de memória. Atualmente, o grid computacional em questão conta com oito nós para processamento e dois nós para controle e gerência do ambiente. Dos nós exclusivos para processamento, dois têm 1TB e 2TB de memória RAM e, respectivamente, 48 e 128 núcleos *hiper-threading* para processamento. Esta última máquina pode escalar até 2048 núcleos e 64TB de memória RAM, constituindo um exemplo expressivo de escalabilidade vertical. Os demais nós têm 512GB de memória RAM e 64 núcleos para processamento cada.

[2] Petabyte equipavale a 10^{15} *bytes*.

[3] *Cluster* e Grids se referem a conjunto de computadores que trabalham juntos na solução de um problema. Usa-se o termo *cluster* para conjuntos fisicamente próximos e normalmente homogêneos e grids para conjunto de *clusters*.

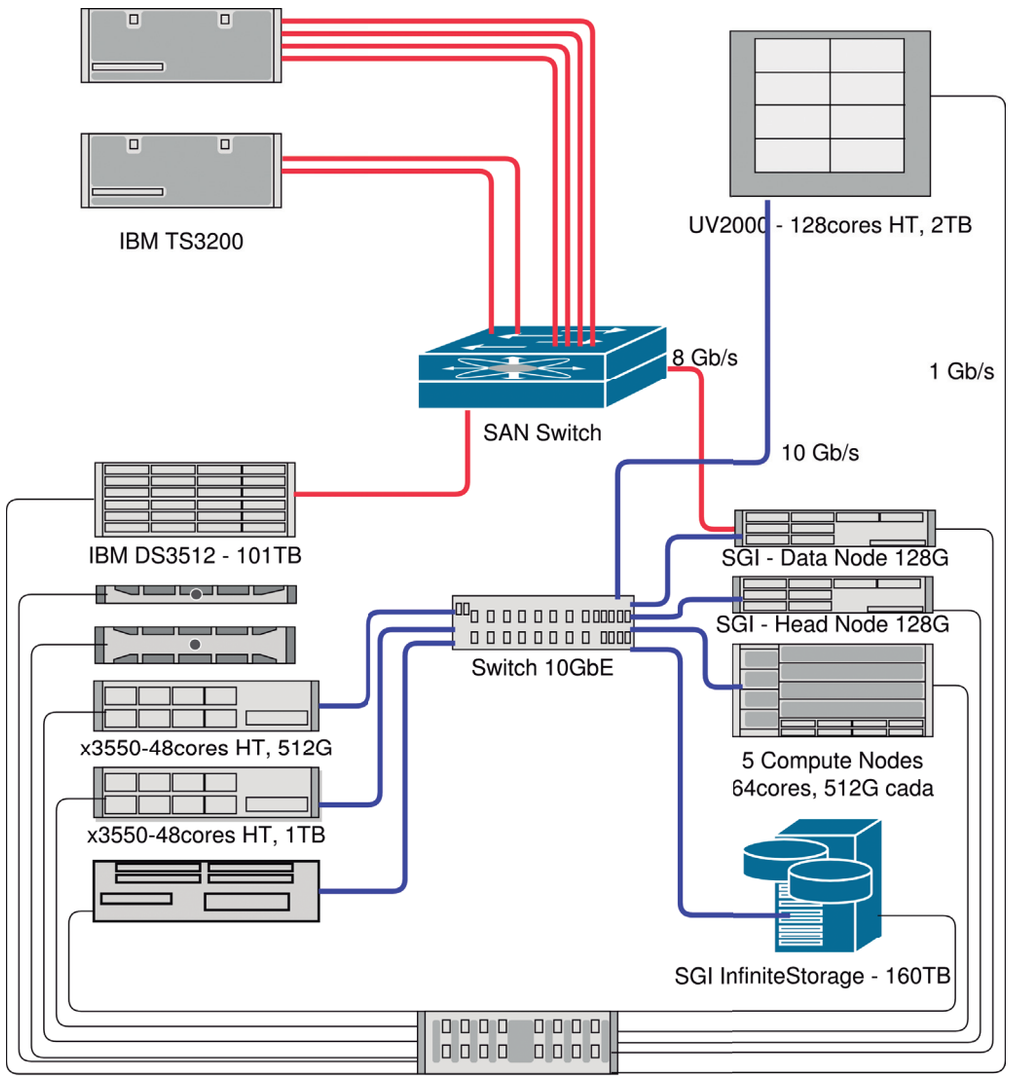

Figura 1. Arquitetura com escalabilidade vertical e horizontal na Embrapa Informática Agropecuária. O grid ilustrado pode ter sua capacidade computacional incrementada adicionando-se novos nós (escalabilidade horizontal), ou expandindo os recursos de alguns nós específicos (escalabilidade vertical).

O ambiente é integrado por duas redes, uma baseada na tecnologia 10 GbE para o tráfego de dados a serem processados e outra, baseada na tecnologia *Gigabit Ethernet*[4], para a gerência do ambiente, por onde trafegam apenas dados de controle dos sistemas. Este grau de conectividade é importante para que o sistema tenha escalabilidade horizontal. Todos os servidores rodam o sistema operacional Linux, uma vez que a maioria das ferramentas de bioinformática disponíveis executam nesta plataforma.

O acesso do usuário ocorre via login remoto (*ssh* - *secure shell*), quando o mesmo necessita executar comandos *shell* no ambiente; ou via interface web, quando utiliza a ferramenta *Galaxy* para executar pipelines de análises. Independentemente dos meios de interação do usuário com o sistema, é vital que a mesma ocorra de forma transparente, sem a obrigatoriedade de se saber muitos detalhes para se utilizar adequadamente o ambiente. Neste sentido, é muito importante em ambientes de processamento científico o uso de sistemas de gerenciamento de recursos (*Distributed Resource Manager System* -DRMS), os quais controlam o escalonamento automático das tarefas (*jobs*) nas máquinas e gerenciam sua execução e resultados. Estes sistemas criam uma abstração do ambiente baseada no conceito de filas de execução. Desta forma, os usuários passam a trabalhar com filas com características específicas, nas quais podem submeter seus trabalhos. A seleção das máquinas, e o instante em que estes trabalhos iniciarão sua execução, ficará a cargo do DRMS determinar.

Em resumo, a infraestrutura apresentada atende atualmente às necessidades de processamento de dados biológicos da Embrapa, e tem a possibilidade de expandir-se de forma bastante efetiva. A capacidade de armazenamento pode ser ampliada, tanto adicionando-se novos módulos ao *storage* já existente quanto adicionando-se novos *storages* ao ambiente. A capacidade de processamento pode ser ampliada com a adição de novos nós, e a capacidade de tratar problemas que exijam grandes quantidades de memória RAM pode ser ampliada com a adição de nós especiais, que tenham a memória necessária disponível.

2.2 *Clusters Hadoop*

O *Hadoop* (BORTHAKUR, 2007; LAM, 2011) é um arcabouço livre de software, mantido pela Apache, voltado para processamento distribuído de grandes quantidades de dados. Além de ser altamente escalável, permitindo a computação distribuída em milhares de computadores, o arcabouço implementa tolerância a falhas no nível da aplicação, provendo serviços de alta disponibilidade. Dentre os diversos componentes do arcabouço, destaca-se sua implementação do modelo de programação *MapReduce* (DEAN, 2008), originalmente proposto pela Google, para processamento e geração de grandes quantidades de dados.

O *MapReduce* aplica a técnica de dividir para conquistar, no qual o problema é dividido em problemas menores e processados separadamente, de forma distribuída, nas máquinas do *cluster*. Neste modelo de programação, a computação é realizada em termos de dois tipos de tarefas *maps* (mapeadores) e *reducers* (redutores). As tarefas de mapeamento são responsáveis por executar a computação sobre frações dos dados de entradas (*splits*) e geram resultados intermediários. Cada *split* é processado por uma instância do mapeador. Os resultados dos mapeadores são consolidados pelos redutores, que agregam os resultados intermediários e geram o resultado final.

O formato dos *splits* variam conforme a natureza dos dados de entrada. Um *split* pode ser, por exemplo, uma quantidade específica de linhas de um arquivo muito grande, uma quantidade fixa de bytes ou um arquivo inteiro, quando os dados de entrada são uma grande coleção de arquivos.

O software que implementa o *MapReduce* (ex. *Hadoop*) cuida do escalonamento das tarefas, ou seja, paraleliza automaticamente a computação dos mapeadores e redutores nas máquinas do *cluster*. As tarefas são monitoradas e, no caso de falhas, são reexecutadas de forma transparente

[4] 10 GbE e Gigabit Ethernet são ambas tecnologias para comunicação em rede de computadores. A diferença principal entre ambas está no fato da primeira possibilitar velocidades de 10 Gbps e a segunda, 1 Gbps.

ao usuário. Estas características facilitam o desenvolvimento das aplicações, pois os programadores não precisam se preocupar com aspectos de escalonamento e tolerância a falhas.

No Laboratório de Modelagem Agroambiental (LMA) da Embrapa Informática Agropecuária, o *Hadoop MapReduce* é utilizado na simulação de cenários agrícolas futuros (NAKAI, 2013). Estas simulações são realizadas por meio de uma metodologia de zoneamento de risco climático utilizando dados de projeções climáticas.

A metodologia de zoneamento utilizada é baseada no cálculo do balanço hídrico, que requer parâmetros, como: coeficiente das culturas, capacidade de armazenamento de água do solo, evapotranspiração e séries temporais de chuva. Utiliza-se um modelo baseado em balanço hídrico chamado Bipzon (ASSAD, 1986; FOSTER, 1984; VAKSMANN, 2000) que calcula o Índice de Satisfação da Necessidade de Água (Isna) para cada coordenada desejada. Este índice é utilizado pelos especialistas para determinar os cenários agrícolas com base no risco climático. Atualmente, os cenários agrícolas futuros são simulados com base em dados históricos de milhares de estações climáticas e dados de modelos de projeção climática.

A Figura 2 ilustra como o *MapReduce* é utilizado na simulação de cenário agrícolas (NAKAI, 2013). Em um primeiro momento, dados climáticos são divididos em *splits* de forma que cada split contenha os dados de uma estação climatológica. O *Hadoop* cria uma instância de mapeador para cada *split*. O mapeador implementa o modelo Bipzon para calcular o valor do Isna correspondente a uma estação para todos os decêndios do ano. Todos esses Isnas ão agregados pelo redutor, que gera um arquivo contendo os Isnas de todas as estações por decêndio. Posteriormente, esses arquivos serão espacializados para criação dos mapas dos cenários agrícolas decendiais.

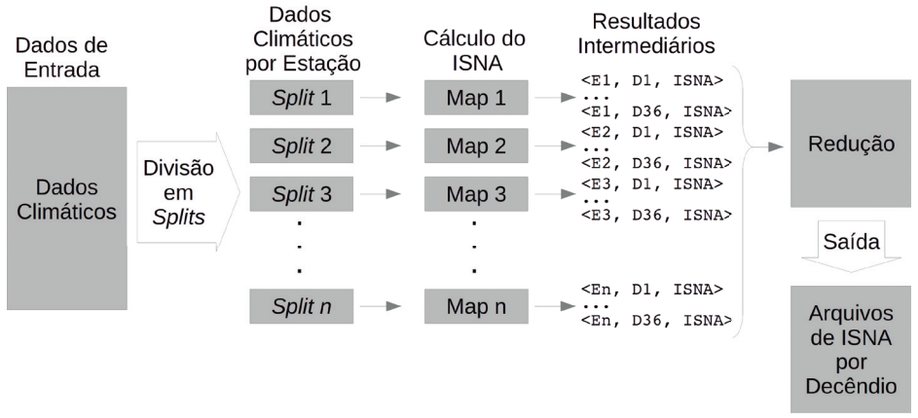

Figura 2. Esquema da simulação de cenários agrícolas utilizando *MapReduce*; *n* é o número de estações. <E*n*, D*i*, Isna> corresponde ao valor do Isna da estação *n*, no decêndio *i*.

O uso do *Hadoop MapReduce* na simulação de cenários agrícolas futuros é apenas um exemplo de uso desta tecnologia na pesquisa agropecuária. Neste caso, especificamente, o *Hadoop* tem possibilitado a realização de estudos envolvendo um grande número de cenários com mais agilidade.

2.3 *Cloud* para infraestrutura computacional

A computação em nuvem (*cloud computing*) tem se tornado um grande foco de pesquisa nas últimas décadas, sendo até considerada como um novo modelo de se fazer computação. Embora este tipo de desenvolvimento seja impulsionado majoritariamente por questões comerciais, essa nova tecnologia pode ser aplicada em benefício de diversas outras áreas. Exemplo disso é sua aplicação neste novo modelo de processo científico intimamente relacionado com a tecnologia da informação.

A utilização de tecnologia da informação em métodos científicos pode ser, de uma forma simples, entendida em duas partes: a parte lógica que utiliza a implementação de um método, um algoritmo ou um modelo, para executar uma ou mais tarefas, e uma parte física, responsável pela execução da parte lógica. O *National Institute of Standards and Technology* (THE NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY, 2010) define a computação em nuvem como sendo:

> "Um modelo para permitir acesso via rede, conveniente e sob demanda, a um conjunto de recursos computacionais configuráveis e compartilhados (como redes, servidores, armazenamento, aplicações e serviços) que podem ser rapidamente provisionados e liberados com o mínimo de esforço de gerenciamento e interação com o provedor do serviço" (tradução nossa).

Ou seja, a computação em nuvem pode ser entendida como uma nova maneira de provisionar recursos computacionais. Seu uso no gerenciamento de infraestrutura altera, diretamente, o modo até então utilizado para realização de atividades de gerência de tecnologia da informação. O controle no uso de ativos computacionais, provisionamento de infraestrutura de hardware (processamento e armazenamento), criação e liberação de serviços, gerenciamento de máquinas virtuais e capacidade de aumentar ou diminuir recursos computacionais, tudo sob demanda, são algumas dessas atividades cuja gerência torna-se extremamente versátil. Esta versatilidade significa otimização no tempo do processo científico.

Além disso, existem outras variáveis envolvidas no uso de nuvens computacionais, principalmente quando utilizada no suporte ao desenvolvimento de pesquisa científica. Atividades de pesquisa envolvem detalhes como propriedade intelectual, sigilo de informações, armazenamento de dados sensíveis, dentre outros relacionados. A maior parte da utilização de nuvens computacionais é dentro de um mercado de venda de serviços, por terceiros, impactando diretamente nesses detalhes. Neste contexto, torna-se mais compreensível as duas vertentes de nuvens computacionais mais utilizadas: nuvens públicas e nuvens privadas. Nuvens públicas são ofertas de serviços sob demanda por empresas, capazes de hospedar infraestrutura computacional, dados e até mesmo software. Já as nuvens privadas são estruturas semelhantes, com os mesmos objetivos, porém estabelecida dentro das próprias instituições que farão uso das mesmas. Além da versatilidade, tem-se o controle total sobre os ativos computacionais, e, principalmente, sobre os dados trabalhados, o que é muito importante em processos científicos.

Existem atualmente alguns pacotes de software que permitem a implementação de ambientes de nuvem, tanto para armazenamento quanto para processamento. O *OpenStack* é o mais difundido, sendo um conjunto de softwares, de código aberto, que permite a criação de uma nuvem para processamento e armazenamento. Uma vez estabelecida, esta nuvem cria um ambiente virtual para pesquisadores executarem seus processos que necessitem de recursos computacionais. Neste contexto, as nuvens têm desempenhado o papel de aproximar a tecnologia da informação

de pesquisadores das diversas áreas do conhecimento. A obtenção de recursos computacionais para o desenvolvimento de suas atividades passa a ser independente de administradores de redes e infraestrutura. As nuvens tem contribuído para tornar a tecnologia da informação um recurso mais imediato no desenvolvimento de métodos científicos. Em um exemplo prático, uma infraestrutura de nuvem permite a um pesquisador alocar uma determinada quantidade de recursos de armazenamento, inserir seus dados, realizar o processamento e armazenar os resultados de forma rápida e independente.

Utilizando o antigo paradigma de compra de equipamentos de TI para a hospedagem de sistemas e execução de tarefas, o custo para o desenvolvimento de certos processos científicos seriam ainda bem altos. As nuvens computacionais, além de todas as características citadas, inovam também em relação a tarifação. Em tempos passados um contrato de prestação de serviços de tecnologia da informação quase sempre reservava os recursos contratados, mesmo se eles não fossem usados. As nuvens permitem o que se chama de elasticidade, de modo que as capacidades dos recursos podem ser aumentadas ou diminuídas facilmente. Esta característica permite um tipo de cobrança sob demanda, ou seja, o usuário paga pelo que for usado. Despesas antes consideradas como Capex (***CAP****ital****EX****penditure*) passam a ser consideradas Opex (***OP****erational****EX****penditure*), otimizando o uso de recursos financeiros.

Em virtude destas características positivas, a Embrapa Informática Agropecuária vem trabalhado com uma nuvem privada destinada especificamente para aplicações em problemas que surgem em projetos de pesquisa voltados para a agricultura.

3 Considerações finais

Não restam dúvidas de que a tecnologia da informação tem impactado e impactará positivamente o processo científico. Sendo assim, o papel da infraestrutura computacional empregada nos projetos de pesquisa torna-se relevante, o que exige uma atenção especial das organizações com relação à infraestrutura disponibilizada. As discussões de aplicações, métodos e técnicas apresentados neste livro mostram a necessidade de uma infraestrutura computacional moderna e adaptável aos mais diversos cenários.

4 Referências

ASSAD, E. D. **Simulation de l'irrigation et du drainage pour les cultures pluviales de riz et de mas en sols de bas-fonds braslia**. Montpellier: IRAT, 1986. 10 p. IRAT. Memories et Travaux, 13.

BORTHAKUR. D. **The Hadoop Distributed File System**: architecture and design. [S. l.]: The Apache Software Foundation, 2007. 14 p. Disponível em: <http://hadoop.apache.org/common/docs/r0.18.0/hdfs_design.pdf>. Acesso em: 18 out 2013.

DEAN, J.; GHEMAWAT S.; Mapreduce: simplified data processing on large clusters, **Commununications of the ACM**, v. 51, n. 1, p. 107-113, 2008.

FOSTER, F. **Simulation du bilan hydrique des cultures pluviales. prsentation et utilisation du logiciel BIP**. Montpellier: IRAT-CIRAD, 1984. 63 p.

HEY, T.; TANSLEY, S.; TOLLE, K. (Org.). **O quarto paradigma**: descobertas científicas na era da eScience. São Paulo: Oficina de textos, 2011. 263 p. il.

LAM, C. **Hadoop in action**. Stamford: Manning, 2011. 312 p. il.

NAKAI, A. M. **Otimizando o Hadoop MapReduce para tarefas pequenas**: um estudo envolvendo simulações de cenários agrícolas. Campinas: Embrapa Informática Agropecuária, 2013. 5 p. (Embrapa Informática Agropecuária. Comunicado técnico, 115).

THE NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY. **NIST Cloud Computing Program**. 2010. Disponível em: <http://www.nist.gov/itl/cloud/>. Acesso em: 1 out. 2014.

VAKSMANN, M. **Le modle bipode**: Logiciel. Bamako: IRAT, 1990.