# Direcionadores tecnológicos da Pesquisa Agropecuária Intensiva em Dados: mapeamento de competências, ferramentas e infraestrutura

*Martha Delphino Bambini*[1], *Roberto Hiroshi Higa* [2], *Maria Beatriz Machado Bonacelli*[3]

[1] Doutoranda no PPG-PCT/IG/Unicamp, Analista de Transferência de Tecnologia na Embrapa Informática Agropecuária, Campinas, SP, martha.bambini@embrapa.br

[2] Pesquisador na Embrapa Informática Agropecuária, Campinas,SP, roberto.higa@embrapa.br

[3] Professora no Departamento de Política, Científica e Tecnológica, do Instituto de Geociências, da Universidade Estadual de Campinas, Campinas, SP, bia@ige.unicamp.br

**RESUMO**

Este artigo apresenta um estudo de monitoramento envolvendo as tendências tecnológicas associadas à aplicação de Pesquisa Computacional Intensiva em Dados em problemas do setor agropecuário. Este estudo foi conduzido a fim de oferecer informações qualificadas para apoio à gestão de uma infraestrutura computacional corporativa com foco em armazenamento e processamento de grandes volumes de dados da Empresa Brasileira de Pesquisa Agropecuária (Embrapa). A infraestrutura visa apoiar a geração de conhecimento de impacto para o domínio agropecuário. O principal resultado desta pesquisa é a descrição dos direcionadores tecnológicos relacionadas à aplicação da Ciência de Dados (*Data Science*) ao setor agropecuário, com foco nas seguintes dimensões: aspectos conceituais, competências e drivers tecnológicos envolvendo o paradigma (ferramentas de software; organização de bases de dados; plataformas especializadas; serviços computacionais). O estudo oferece informações tecnológicas qualificadas para apoiar a gestão da infraestrutura computacional corporativa da Embrapa.

**PALAVRAS-CHAVE:** Pesquisa Computacional; Agropecuária; Ciência de Dados; Big Data.

**ABSTRACT**

This article presents a monitoring study involving the technological trends associated to the application of Data Intensive Computational Research to agricultural sector problems. This study was conducted in order to provide qualified information to manage a corporate computing infrastructure focused on the storage and processing of large volumes of data of the Brazilian Agricultural Research Corporation (Embrapa). It would support the generation

of impacting knowledge related to Agriculture. The main result of this research is the description of the technological drivers related to the application of Data Science to the agricultural sector, focusing on the following dimensions: conceptual aspects, competences and technological drivers involving the paradigm (software tools; Databases, specialized platforms, computer services). The study offers qualified technological information to support the management of Embrapa´s corporate computing infrastructure.

**KEYWORDS:** Computing Research; Agriculture; Data Science; Big Data.

## INTRODUÇÃO

Vive-se hoje em um momento marcado pela 3ª fase da Revolução das Tecnologias de Informação e Comunicação (TIC), descrita por Cowhey, Aronson e Abelson (2009). Os autores apontam que esta fase é marcada por um ponto de inflexão, no qual as novas arquiteturas de TIC – chamadas de Nuvem ou Grid – estruturam-se em um conjunto bastante distintos de capacidades e organização de mercado do que nas fases anteriores. Estas arquiteturas consideram três tendências relacionadas: (i) a existência de redes de Internet de banda larga disponíveis, a custo acessível; (ii) a integração de um armazenamento massivo e barato de informação, associada a uma nova arquitetura de redes e serviços; e (iii) a emergência de sistemas de computação virtual que controlam e se utilizam de muitos computadores, incluindo supercomputadores, de forma a atender demandas dos usuários e aumentar eficiência e produtividade.

Avent (2014) destaca que esta nova onda tecnológica é formada por um conjunto de novas tecnologias – como inteligência de máquina, Internet Ubíqua (onipresente) e robótica avançada – capazes de possibilitar o desenvolvimento de várias invenções, entre ela: veículos não tripulados, drones, sistemas automáticos de tradução de idiomas, tecnologias móveis que possibilitam diagnóstico e tratamento médico à distância e também aprendizado virtual. No contexto desta nova fase da revolução das TIC, o volume de geração de dados de pesquisas em áreas de interesse agropecuário vem crescendo rapidamente e as necessidades de armazenamento de processamento de grandes volumes de dados se tornam prementes.

Bell (2009) descreve as 3 atividades básicas da computação intensiva em dados: captura, curadoria e análise. Os dados são coletados em escalas e formatos variados, podendo envolver resultados de experimentos e de observações individuais, laboratórios únicos, grupos de laboratórios parceiros, redes de laboratórios internacionais em larga escala bem como conjuntos de dados pessoais de indivíduos pesquisados. No Século XXI, a grande maioria do

grande volume de dados científicos é capturada por instrumentos, de forma contínua, ou informação geradas a partir de modelos computacionais (HEY, TANSLEY e TOLLE, 2009), processados por softwares específicos, sendo as informações resultantes armazenadas em bases de dados. Estes dados podem ser armazenados por longo prazo e disponibilizados de forma pública, curados e gerenciados de forma a garantir sua análise continuada, embasando a geração de novas teorias científicas derivadas. Neste contexto, a velocidade de geração de dados pode ser muito grande, dificultando o processo de captura e armazenamento. Assim, há que se criar um conjunto genérico de ferramentas que cubram o conjunto completo destas atividades científicas: captura, validação de dados, curadoria, análise e armazenamento.

O objetivo deste trabalho é identificar e analisar os principais direcionadores tecnológicos relacionados ao conceito de Pesquisa Computacional Intensiva em Dados - chamada no documento de *Data Science* - aplicada ao setor agropecuário. Este estudo se insere no âmbito das atividades do projeto GDAE, intitulado "Estruturação de processos e redes de relacionamento em apoio à gestão do arranjo Armazenamento e Processamento de Dados Experimentais da Embrapa - DATAExp". Este projeto tem entre seus objetivos delinear os processos de estruturação e gestão de uma nova infraestrutura computacional institucional para o uso por projetos de pesquisa da Empresa Brasileira de Pesquisa Agropecuária (Embrapa) que demandem armazenamento e processamento de grandes volumes de dados para gerar conhecimento de impacto para o domínio agropecuário. Algumas temáticas de pesquisa consideradas relacionadas ao uso desta plataforma são: genética e genômica, agricultura de precisão e modelagem climática. A adoção de uma infraestrutura computacional institucional visa ao uso racional dos recursos computacionais da empresa e à otimização dos investimentos em novos equipamentos e infraestruturas de processamento de alto desempenho.

## MATERIAL E MÉTODOS

O trabalho aqui apresentado consiste em um estudo de monitoramento tecnológico, conforme descrito em Porter et al. (1991), envolvendo a aplicação de Pesquisa Computacional Intensiva em Dados ao domínio agropecuário. O estudo foi conduzido a partir de análise de literatura de forma a mapear as principais tendências tecnológicas relacionadas ao tema. A busca foi conduzida na base da *Web of Science* a partir das palavras-chave "*Data Science*" e "Agriculture", sendo complementada com consulta no *Google Scholar* e na plataforma LinkedIn. Foram consideradas as seguintes dimensões: aspectos conceituais, competências da

equipe, e drivers tecnológicos envolvendo o paradigma (ferramentas de software; organização de bases de dados; plataformas especializadas; serviços computacionais).

## RESULTADOS E DISCUSSÃO

Esta seção apresenta os conceitos de base relacionados à pesquisa conduzida e o levantamento de competências relacionadas à Ciência de Dados; e um mapeamento dos Drivers Tecnológicos envolvendo infraestrutura, produtos e serviços relacionados à aplicação deste paradigma ao setor agropecuário.

### *Análise da tecnologia e seu contexto: aspectos conceituais*

Jim Gray (HEY, TANSLEY e TOLLE, 2009) descreve o 4º paradigma da ciência marcado por uma dinâmica de pesquisa colaborativa, em rede e direcionada a dados, chamada pelo autor de *e-Science.* O termo descreve a síntese de Tecnologia de Informação e Ciência, levando a grandes mudanças e transformações nas atividades de pesquisa científica.

A ciência intensiva em dados, ou *Data Science*, se refere a um campo de estudo interdisciplinar que se utiliza de processos científicos e computacionais para extrair conhecimento, padrões, tendências e insights a partir de conjuntos de dados de vários formatos, estruturados ou não-estruturados. Bell, Hey e Szalay (2009) descrevem o "dilúvio de dados" que vem ocorrendo nas últimas décadas, quando dados oriundos de sensores, de simulações ou outras ferramentas computacionais de análise necessitam ser analisados em um processo científico intensivo em dados. Os autores destacam que, mesmo com um enorme potencial identificado, a ciência intensiva em dados tem se desenvolvido ainda de forma lenta, talvez pela falta de entendimento da comunidade científica em relação a este tema e às ferramentas práticas para criar, gerenciar e utilizar estas bases de dados. Ressaltava-se, naquele momento, a urgência por novas ferramentas e tecnologias para tornar possível os trabalhos de *Data Science*. Alguns campos científicos de aplicação citados pelos autores são: a Astronomia, Física de alta energia e a Genômica.

### *Fenômeno Big Data*

Outro conceito que também se relaciona com a *Data Science*, que se transformou em uma *buzzword* na área de negócios, é *Big Data*. Conforme Diebold (2012), o termo *Big Data* - relacionado à Ciência da Computação, à Estatística e à Econometria - se originou de trabalhos não acadêmicos da empresa Silicon Graphics Inc. (SGI) em meados da década de 1990. O autor destaca que não é possível dizer que seja um novo campo de estudos tendo em vista a existência de trabalhos prévios já desenvolvidos em Ciência da Computação e Estatística já há

muitos anos. Mas, as técnicas e ferramentas utilizadas para lidar com a questão de *Big Data* são realmente inovadoras, como C*loud computing*, associação de algoritmos "massivamente paralelos", métodos para controlar descobertas falsas ao testar milhões de hipóteses, entre outras. Pospiech e Felden (2012) indicam que o termo *Big Data* vem sendo cada vez mais estudado em publicações científicas, em várias disciplinas e sob várias definições. No que se refere ao mercado, Few (2012) destaca que este termo vem sendo veiculado por várias publicações na área de *business intelligence* (BI), configurando-se pura e simplesmente como uma campanha de marketing de empresas relacionadas ao setor.

O fenômeno "*Big Data*" é geralmente caracterizado na literatura por 3 Vs: **volume** de dados (quantidade); **velocidade** de aquisição (criação, acumulação, inserção e processamento dos dados) e grande **variedade** de tipos de dados, tornando o processo de gestão e análise de dados mais complexo e pesado. Dados mais tradicionais são, em geral, estruturados, como caracteres, números, etc. Nas últimas décadas, as aplicações e sistemas passaram a gerar um grande aumento de dados não-estruturados como textos, áudio, vídeo, imagens, dados geoespaciais e de internet (como logs e registros de cliques). As implicações desta grande variedade de tipos de dados são: maior complexidade na construção e inserção destes dados em uma base computacional e maior dificuldades na transformação de dados para processamento analítico e computacional. Na próxima subseção, serão apresentados as competências e o perfil do Cientista de Dados.

***Novas competências e um novo perfil profissional para lidar com Data science***

Minelli, Chambers e Dhiraj (2013) destacam que os grandes conjuntos de dados não representam o maior desafio no contexto estudado, mas sim, a identificação e o desenvolvimento de métodos menos custosos e confiáveis para extrair valor dos conjuntos de *terabytes* e *petabytes* de dados. Além de empregar tecnologias e novos métodos de análise, com maior velocidade e escala para lidar com alta complexidade, torna-se necessário também **contratar e/ou capacitar profissionais e criar novos processos de trabalho**.

Davenport e Patil (2012) descrevem um **novo perfil profissional,** o cientista de dados ou *data scientist,* termo cunhado em 2008: Este profissional deve ter o treinamento e as habilidades necessárias para fazer descobertas a partir de um grande conjunto de dados. A demanda por estes profissionais é crescente, mas a oferta nem tanto, tendo em vista que não existem programas de graduação e/ou pós-graduação com este grau de especificidade.

Existem hoje alguns programas específicos de formação e iniciativas de capacitação *on the job* por parte de algumas empresas.

Uma consulta efetuada em 27/01/2017 na seção de empregos (*Jobs*) da plataforma LinkedIn (2017) utilizando o termo "*data scientist*" retornou 58 empregos no Brasil com este termo. Já nos Estados Unidos, são 11.776 vagas anunciadas para "data scientists". Na Inglaterra são 1.354 vagas para este cargo; na Alemanha, 658 vagas; na França são 449; na Holanda 229; na Espanha são 159. Em países da América Latina temos 49 vagas no México; 11 vagas no Chile e 10 na Argentina. Ainda no LinkedIn, o termo "data scientist" retornou 63.203 pessoas com este termo em seu perfil e 110 grupos relacionados ao tema.

Existe ainda pouco consenso sobre o papel deste tipo de profissional em uma organização, sua alocação funcional; como podem agregar valor e como sua performance deve ser mensurada (DAVENPORT e PATIL, 2012). Em termos de competências técnico-científicas, espera-se que cientistas de dados possuam fundamentos sólidos em matemática, estatística, probabilidade e computação, sendo a formação científica desejável independente do campo de formação. Habilidades específicas podem ser construídas em treinamentos *on-the-job* e eventuais cursos sobre ferramentas tecnológicas. Os profissionais devem saber codificar software, ainda que não seja sua principal habilidade, e também conseguir comunicar os resultados de seu trabalho de forma a que os *stakeholders* consigam compreendê-los e utilizá-los. É muito importante que cientistas de dados se conectem com comunidades de prática, seja na organização em que trabalha ou externamente com outros profissionais da área. Outras habilidades indicadas pelos autores são: saber identificação e integração de bases de para completar dados faltantes e gerar melhores resultados; extrair *insights* e informações relevantes; comunicar estes *insights* e informações às equipes; apresentação e visualização criativa dos dados, criando padrões de interesse; sugerir implicações para o direcionamento dos negócios, sobre processos executados e produtos comercializados.

***Análise de Drivers Tecnológicos envolvendo Data Science***

Segundo Minelli, Chambers e Dhiraj (2013), as crescentes demandas para tratamento de dados com **volume, variedade e velocidade** aumentam as demandas por plataformas de computação e tecnologias de software capazes de lidar com escala, complexidade e velocidade a fim de garantir a competitividade de organizações, sejam empresas ou institutos de pesquisa. Os autores indicam que diferentes sistemas como software para gestão e análise

de dados, sistemas livres e proprietários bem como configurações de hardware, estão sendo combinadas para criar novas soluções em Tecnologias de Informação para lidar com os desafios de *Big Data Analytics.*

Davenport e Patil (2012) destacam também que é comum encontrar casos em que os cientistas de dados desenvolvem suas próprias ferramentas, conduzindo pesquisas em estilo acadêmico, muito embora Minelli, Chambers e Dhiraj (2013) apontem a existência também soluções proprietárias.

Pospiech e Felden (2012) indicam que os pesquisadores divergem quanto ao uso debasesescalabilidade e performance exigida por este tipo de aplicação. Outras abordagens mencionadas pelos autores são: bases de dados funcionando em paralelo; NoSQL; e bases de dados distribuídas para analisar conjuntos de dados muito grandes.

As arquiteturas computacionais para processamento intensivo em dados precisam ser escaláveis, ou seja, devem ser projetadas para que sua capacidade de processamento e armazenamento acompanhe o crescimento da complexidade dos problemas tratados e o correspondente volume de processamento de dados requerido. Usualmente, essas arquiteturas são classificadas em duas categorias, do ponto de vista de escalabilidade (CINTRA, NAKAI e CORREA, 2014 ): **escalabilidade vertical** em que os computadores podem ter até milhares de processadores, dezenas de *terabytes* de memória RAM e sistemas de armazenamento de dados especializados que podem atingir *petabytes* de capacidade; e **escalabilidade horizontal**, representadas por *cluster* e *grids* de computadores, cuja capacidade é expandida pela adição de novos nós, ou seja, computadores ao sistema. Enquanto este último possui a vantagem de ter um custo por processador ou memória mais acessível, é preciso ter em mente que existem problemas para os quais uma solução só é viável em sistemas verticalmente escaláveis. É por esse motivo que centros de computação científica, dedicados a apoiar projetos de pesquisa intensivo em dados (ex: *Commonwealth Scientific and Industrial Research Organisation* (CSIRO, 2017)) em geral possuem os dois tipos de soluções.

O projeto Apache Hadoop (APACHE, 2017), formado por uma biblioteca de software livre, detecta e manipula falhas na camada de aplicação, provendo um serviço de alta disponibilidade sobre o cluster de computadores, cada um dos quais pode estar sujeito a falhas. Ele é composto por diversos módulos, dentre os quais destaca-se o MapReduce (DEAN e GHEMAWAT, 2008), um modelo de programação simples, originalmente desenvolvido pelo Google, e que permite o processamento e geração de grandes volumes de

dados. Ele baseia-se no princípio de solução de problemas conhecido como "dividir para conquistar", consistindo em dividir o problema original em diversos problemas menores que podem ser distribuídos para serem processados, de forma independente e em paralelo, em nós de um cluster de computadores; e então, posteriormente, consolidados para obter a solução do problema original. Além dos módulos que o compõem, o Hadoop também possui um rico ecossistema de projetos associados, com funcionalidades para processamento de grandes volumes de dados, que podem ser consultados em Apache (2017).

Um outro direcionador-chave do fenômeno *Big Data* e das pesquisas em *Data Science* é o *Cloud computing*, ou computação em nuvem, que pode ser vista com uma forma de consumo de recursos computacionais (máquina virtual, *storage* ou aplicação) como um serviço, sem a necessidade de constituir e manter uma infraestrutura computacional própria. Esse tipo de serviço possui vários atrativos para o usuário, tais como o auto provisionamento, a elasticidade e o pagamento apenas pelo uso dos recursos; pode ser oferecido por provedores comerciais (nuvem pública), pela própria instituição (nuvem privada) ou em um modelo híbrido, que contempla os dois casos anteriores; sendo o serviço categorizado de 3 diferentes formas: infraestrutura como serviço (IaaS), plataforma como serviço (PaaS) e software como serviço (SaaS). Esses modelos subsidiam novos modelos de negócios envolvendo a prestação de serviços associados ao armazenamento, organização, validação, processamento e análise de grandes volumes de dados.

Embora, seu uso esteja bastante difundido em aplicações comerciais, a comunidade de computação de alta performance ainda tem hesitado em entrar nesse mercado (SCIENTIFIC COMPUTING WORLD, 2017). Isso, contudo, parece estar mudando à medida que caem os custos e tornam-se disponíveis novos modelos de programação que permitem a execução mais eficiente na nuvem de processamentos de tarefas que exigem alta performance. Provedores de nuvens públicas estão instalando configurações de hardware próprios para HPC (*High Perfomance Computing*) ao mesmo tempo em que nuvens privadas estão dando aos usuários de computação científica a experiência sobre como executar tarefas em um ambiente em nuvem. Entretanto, ainda na mesma referência, o autor chama atenção para o fato de que nem toda aplicação de HPC é apropriada para execução em um ambiente baseado em nuvem. Tarefas fracamente acopladas sem muitas operações de entrada e saída (I/O) são bons candidatos para execução em ambiente de nuvem; enquanto aplicações com requerimentos de

memória muito elevados ou tarefas altamente paralelizadas em dezenas de milhares de núcleos não.

Ainda, outras questões relacionadas ao uso dessa tecnologia são a **segurança da informação e o armazenamento de dados sensíveis** (CINTRA, NAKAI e CORREA, 2014), quando se está utilizando o serviço oferecido por provedores públicos. Uma alternativa para essa situação é a estruturação de uma nuvem híbrida e a utilização de uma estratégia que preserve os dados mais sensíveis na parte privada da nuvem. *Scientific Computing World* (2017) aponta que a tendência atual é que o próximo passo no desenvolvimento da tecnologia de nuvem seja a interligação de várias nuvens privadas, formando uma federação de nuvens. Dessa forma, uma empresa multinacional (ou um conjunto de instituições) com centros de dados em múltiplas localizações poderá balancear a carga de processamento entre todos os centros de dados. Para instituições de pesquisa, essa é uma tendência que vai ao encontro com o que se espera para o futuro da ciência intensiva em dados e e-Science (HEY, TANSLEY e TOLLE,2009).

## CONCLUSÕES

Esta pesquisa buscou contribuir para o projeto GDAE da Embrapa gerando informações qualificadas para apoiar a estruturação dos processos e as redes de relacionamento para apoiar a gestão desta nova infraestrutura computacional. Entende-se que, em consonância com as tendências encontradas na literatura e com o trabalho de ARS (2013), as instituições de pesquisa que, como a Embrapa, busquem uma posição de liderança em computação científica aplicada a agropecuária, necessitam efetuar investimentos em infraestrutura computacional, em novas práticas e no desenvolvimento de competências. Os investimentos devem prever: aumento da eficiência computacional na instituição com a redução de redundância nas compras de hardware; desenvolver a capacidade de compartilhar dados entre laboratórios, instituições e disciplinas; aprimorar competências da força de trabalho; criação de novos métodos computacionais e estabelecimento de padrões para coleta de dados, estruturação, bem como sua documentação e disseminação, entre outros.

## REFERÊNCIAS

APACHE - *Apache Software Foundation*. Disponível em: <https://www.apache.org/>. Acesso em: 06 jan.2017.

AVENT, R. The third great wave. *The Economist*, October, v. 4, 2014.

ARS - Agricultural Research Service. United States Department of Agriculture – USDA. *Big Data and Computing Building a Vision for ARS Information Management. Workshop Summary Feb. 5-6, 2013*. Disponível em: <https://www.ars.usda.gov/> Acesso em: 13.fev.2017.

BELL, G. Foreword. In: HEY, T., TANSLEY, S. , TOLLE, K. M. *The fourth paradigm: data-intensive scientific discovery*. Redmond, WA: Microsoft Research, 2009. pp. xi-xiv.

BELL, G.; HEY, T.; SZALAY, A. Beyond the data deluge. *Science* v. 323, n. 5919, p. 1297-1298, 2009.

CINTRA, L.C; NAKAI, A.; CORREA; J. L. Métodos, procedimentos e técnicas utilizadas na construção de AgroTIC. In Tecnologias da Informação e Comunicação e suas relações com a agricultura. Massruhá, S. M. F. S.; Leite, M. A. A.; Luchiari Jr., A.; Romani, L. A. S. (eds). Embrapa. Brasília, DF. 2014.

COWHEY, P.F., ARONSON, J.D., ABELSON, D. *Transforming Global Information and Communication Markets: The Political Economy of Innovation*. Cambridge, USA: Massachusetts Institute of Technology, 2009. 341 p.

CSIRO - Commonwealth Scientific and Industrial Research Organisation. Disponivel em: <https://confluence.csiro.au/display/SC/Facilities> Acesso em: 06 jan.2017.

DAVENPORT, T. H.; PATIL, D. J. Data scientist. *Harvard business review*, v. 90, n. 5, p. 70-76, 2012.

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

DIEBOLD, F. X. "On the Origin(s) and Development of the Term "Big Data". *PIER Working Paper* 12-037. Philadelphia, PA: Penn Institute for Economic Research, 2012.

FEW, S. Big data, big ruse. Visual Business Intelligence Newsletter, 2012.

HEY, T., TANSLEY, S. , TOLLE, K. M. *The fourth paradigm: data-intensive scientific discovery*. Redmond, WA: Microsoft Research, 2009. 252p.

PORTER, A.; BANKS, J; ROPER, A. T.; MASON,T.; ROSSINI, F.; WIEDERHOLT, B. *Forecasting and management of technology*. New York, NY: John Wiley and Sons, 1991. pp. 114-132.

POSPIECH, M.; FELDEN, C. Big data–a state-of-the-art. ( July 29, 2012). *AMCIS 2012 Proceedings*. Paper 22. Disponível em: <http://aisel.aisnet.org/> Acesso em: 24 Jan 2017.

LINKEDIN. Disponível em: < https://pt.linkedin.com/> Acesso em: 06 jan.2017

MINELLI, M.; CHAMBERS, M.; DHIRAJ, A. What Is Big Data and Why Is It Important?. In: *Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses*, 2013. pp. 1-18.

SCIENTIFIC COMPUTING WORLD. *HPC finally climbs into the cloud*. 16 de fevereiro de 2016. Disponível em: <https://www.scientific-computing.com/feature/hpc-finally-climbs-cloud/> Acesso em: 6 Jan.2017.

SZALAY,A. S. , BLAKELEY,J. A. Gray's laws: database-centric computing in science. In: HEY, T., TANSLEY, S. , TOLLE, K. M. *The fourth paradigm: data-intensive scientific discovery*. Redmond, WA: Microsoft Research, 2009. pp. 5-11.