J. Health Inform. 2016 Abril-Junho; 8(2):73-9

Artigo Revisão

JHI JOURNAL OF HEALTH INFORMATICS

Big Data e Nuvens Computacionais: Aplicações em Saúde Pública e Genômica

Big data and cloud computing: Applications in Public Health and Genomics

Big Data y Cloud Computing: Aplicaciones en Salud Pública y Genómica

Fabricio Alves Barbosa da Silva[1]

Descritores: Saúde Pública; Genômica; Banco de Dados; Computação de Alto Desempenho

RESUMO

Big Data é um termo utilizado para descrever conjuntos de dados cuja captura, armazenamento, distribuição e análise requerem métodos e tecnologias avançadas devido a qualquer combinação de seu tamanho (volume), a frequência de atualização (velocidade) e diversidade (heterogeneidade). Este artigo apresenta uma revisão bibliográfica sobre aplicações de Big Data em saúde pública e em genômica. São descritos diversos exemplos, e alguns desafios tecnológicos relacionados com a análise destes dados são identificados. Também é discutida a utilização de nuvens computacionais no processamento de Big Data. No nosso ponto de vista, a nuvem computacional é uma plataforma adequada para o processamento de grandes volumes de dados, e pode ser usada também em diversas aplicações relacionadas à saúde pública e genômica. Diversos trabalhos disponíveis na literatura e citados neste artigo corroboram esta visão.

Keywords: Public Health; Genomics; Database; High Performance Computing

ABSTRACT

Big Data is a term used to describe data sets whose capture, storage, distribution and analysis require advanced methods and technologies due to any combination of its size (volume), the update frequency (speed) and diversity (heterogeneity). This paper presents a literature review of Big Data applications in public health and genomics. Several examples are described, and some technological challenges related to the analysis of these data are identified. The use of computational clouds for Big Data processing is also discussed in this paper. In our view, the cloud is an appropriate platform for the processing of large volumes of data, and it can be used in several applications related to public health and genomics. Several studies available in the literature and cited in this paper corroborate this view.

Descriptores: Salud Pública; Genómica; Base de Datos; Computación de Alto Rendimiento

RESUMEN

Big Data es un término usado para describir conjuntos de datos cuya captura, almacenamiento, distribución y análisis requieren métodos y tecnologías avanzadas, debido a una combinación de su tamaño (volumen), la frecuencia de actualización (velocidad) y la diversidad (heterogeneidad). Este artículo presenta una revisión de la literatura de aplicaciones Big Data en la salud pública y la genómica. Se describen varios ejemplos, y algunos de los desafíos tecnológicos relacionados con el análisis de estos datos se identifica. También discute es el uso de nubes computacionales en el procesamiento de grandes volúmenes de datos. En nuestra opinión, la plataforma de computación en la nube es adecuado para el procesamiento de grandes volúmenes de datos, y también puede ser utilizado en diversas aplicaciones relacionadas con la salud pública y la genómica. Varios estudios disponibles en la literatura y citados en este artículo corroboran esta opinión.

[1] *Pesquisador em Saúde Pública. Programa de Computação Científica – Fundação Oswaldo Cruz. Professor do Corpo Docente Permanente do Programa de Pós-Graduação Strictu Sensu em Biologia Computacional e Sistemas do Instituto Oswaldo Cruz - IOC/FIOCRUZ, Rio de Janeiro (RJ), Brasil.*

Autor Correspondente: **Fabricio Alves Barbosa da Silva**
e-mail: **fabs@fiocruz.br**

Artigo recebido: **07/07/2014**
Aprovado: **19/01/2016**

www.jhi-sbis.saude.ws

INTRODUÇÃO

Em muitas áreas da ciência, avanços tecnológicos e conceituais estão resultando na geração cada vez mais rápida de grandes quantidades de dados. Enquanto a pesquisa científica sempre envolveu a coleta e organização dos dados, o volume, variedade e a velocidade da produção atual de dados científicos apresentam novas oportunidades e desafios, tanto em escala quanto em complexidade.

Várias definições do termo "Big Data" são encontradas na literatura. Por exemplo, podemos definir Big data como um termo usado para descrever conjuntos de dados cuja captura, armazenamento, distribuição e análise requerem métodos e tecnologias avançadas, devido a qualquer combinação de seu tamanho (volume), a frequência de atualização (velocidade) e diversidade (heterogeneidade)[1-2].

O termo "Big Data" está associado a uma nova geração de tecnologias e arquiteturas concebidas para extrair valor de grandes volumes de uma ampla variedade de dados, tornando viável uma alta velocidade de captura e a análise de grandes volumes de dados. Este mundo dos grandes dados exige uma mudança na arquitetura de computação de modo a atender tanto os requisitos de armazenamento dos dados quanto de processamento intensivo, distribuído, necessários para analisar grandes volumes de dados de uma maneira segura. A maior parte dos dados classificados como Big Data é composta de informação não-estruturada e normalmente não é simples analisá-los através de bancos de dados tradicionais. Entretanto, o poder preditivo proveniente da análise de grandes volumes de dados tem sido explorado recentemente em campos como saúde pública, genômica e medicina[2-3].

O conceito de Big Data está fortemente relacionado ao fenômeno de dilúvio de dados (*Data Deluge*)[4]. O dilúvio de dados se refere à situação em que o crescimento exponencial na geração de novos dados torna cada vez mais complexo o seu gerenciamento e análise.

O fenômeno de dilúvio de dados pode ser ilustrado através do crescimento do GenBank[5]. O GenBank é um banco de dados de sequências genéticas mantido pelo NIH (EUA), e consiste em uma coleção anotada de todas as sequências de DNA publicamente disponíveis. Atualmente o Genbank dobra de tamanho a cada 18 meses. Esta tendência vem sendo confirmada em anos anteriores e deve ser mantida nos próximos anos.

Entretanto, cabe ressaltar que o crescimento exponencial de dados não requer apenas novas tecnologias para acessar e integrar estes dados, mas, também, o desenvolvimento de novos métodos de análise que sejam computacionalmente eficientes e eficazes no processamento de dados que podem ser muito ruidosos (e.g. dados cujo coeficiente de determinação - R^2 - em relação a um modelo estatístico linear generalizado é menor que 0.5)[6]. Ou seja, para se extrair o potencial e a promessa do Big Data em saúde pública, a colaboração de pesquisadores de áreas tão diversas como ciência da computação, métodos quantitativos e saúde pública é fundamental. Big Data em Saúde Pública define um tema que é interdisciplinar na sua essência, no qual colaboração é indispensável.

O mesmo impacto potencial está presente na genômica. O sequenciamento de genomas é uma aplicação pioneira de Big Data, afinal um único genoma humano é composto por cerca de 3 bilhões de pares de bases de DNA. Tecnologias de sequenciamento de nova geração permitem que o sequenciamento completo de genomas em larga escala seja viável em termos de custo e tempo. A caracterização em larga escala do genoma humano envolve a geração e interpretação de um enorme volume de dados, em uma escala sem precedentes, e um dos benefícios potenciais é medicina personalizada para pacientes de câncer[7].

Neste artigo são descritos diversos exemplos de aplicação de Big Data e Nuvens Computacionais em saúde pública e genômica, e são identificados alguns desafios tecnológicos relacionados a análise de grandes volumes de dados. No nosso ponto de vista, a nuvem computacional é uma plataforma adequada para o processamento de grandes volumes de dados, e pode ser usada também em diversas aplicações relacionadas à saúde pública e à genômica. Diversos trabalhos disponíveis na literatura e citados neste artigo corroboram esta visão[1-2, 8-10].

Neste artigo é apresentada uma revisão da literatura que relaciona Big Data e nuvens computacionais às áreas de saúde pública e genômica. Diversos exemplos são descritos e alguns desafios tecnológicos relacionados a análise de grandes volumes de dados são identificados. A literatura sobre Big Data e nuvens computacionais em geral é bastante recente, e a quase totalidade dos artigos citados neste trabalho cobrem o período 2008-2015.

Este artigo está organizado como segue: na seção 2 descrevemos diversos exemplos de aplicação de Big Data relacionados à saúde pública. A seção 3 é dedicada à discussão de Big Data em genômica. Na seção 4 aprofundamos a discussão sobre a utilização de nuvens computacionais como plataforma de processamento e armazenamento de Big Data, mantendo o foco em aplicações genômicas e de saúde pública. A seção 5 discorre sobre diversos desafios relacionados à utilização de Big Data em genômica e saúde pública e a seção 6 apresenta nossas considerações finais.

BIG DATA EM SAÚDE PÚBLICA

A revolução das comunicações, através da explosão da telefonia móvel e da internet, teve duas consequências principais: (1) todos os tipos de comunicação modernos são agora digitais e (2) o número de usuários de dispositivos permitindo comunicação digital está na casa dos bilhões, se aproximando rapidamente da cobertura completa em várias partes do mundo. Como consequência, uma fração cada vez maior do que fazemos e dizemos, incluindo comportamentos epidemiologicamente relevantes[11], tais como decidir sobre medidas de prevenção e opções de tratamento, bem como informar sobre sintomas de doenças, é armazenado em formato digital, sendo então acessíveis à análise. Extrair informações significativas a

partir deste imenso volume de dados é um desafio, mas também representa uma enorme oportunidade para a epidemiologia[11]. Devido ao seu alcance e à velocidade de propagação da informação, a exploração deste grande volume de informações, oriundo da digitalização das comunicações, como meio auxiliar aos sistemas de vigilância epidemiológica tradicionais, é uma oportunidade importante para países de dimensões continentais como o Brasil. Entre possíveis aplicações destacam-se exemplos relacionados ao monitoramento epidemiológico, farmacovigilância e mapeamento de risco de doenças transmissíveis, como descrito a seguir.

Monitoramento epidemiológico: o uso de Big Data em conjunto com um grande poder de processamento pode servir como sistema complementar às redes de vigilância epidemiológicas tradicionais. Por exemplo, é possível utilizar logs de busca do Google para rastreamento da gripe em uma população[12]. Ginsberg e colegas mostraram que a frequência relativa de certas consultas no Google está altamente correlacionada com a porcentagem de visitas ao médico nas quais um paciente se apresenta com sintomas de gripe. Um sistema similar foi desenvolvido para o monitoramento da dengue[13]. Outros sistemas equivalentes, desta vez utilizando mensagens do Twitter como fonte de dados para monitorar a gripe, são também descritos na literatura[14]. Cabe ressaltar, entretanto, que estes sistemas estão sujeitos a erros, como superestimações eventuais[15].

Farmacovigilância: O processamento de um grande volume de dados pode ser usado para identificar associações entre drogas e efeitos adversos até então desconhecidas. Por exemplo, White e colegas investigaram a utilização de logs de busca em farmacovigilância[16], processando 82 milhões de logs de buscas, obtidos de 6 milhões de usuários, durante o ano de 2010. Além disso, a combinação de dados de múltiplas fontes pode tornar mais eficaz a detecção de efeitos adversos associados a fármacos. Uma combinação possível é o uso conjunto de notificações de efeitos adversos (*Adverse Event Reporting System* - AERS) e Prontuários eletrônicos (*Electronic Health Records* - EHR). Harpaz e colegas[17] utilizaram esta combinação de dados heterogêneos para identificar uma nova associação entre rasburicase e pancreatite aguda, através do processamento de 4 milhões de AERS e 1,2 milhão de EHRs.

Vários trabalhos recentes investigaram a utilização de mensagens curtas do Twiter em farmacovigilância, através do processamento de milhões de mensagens[18-19]. Além do Twitter, alguns trabalhos que descrevem a utilização de outras mídias sociais em estudos de farmacovigilância, como fóruns online, podem ser encontrados na literatura[20].

Mapeamento dinâmico de risco de doenças transmissíveis: o conceito de Big Data abre perspectivas interessantes para o desenvolvimento de diversos sistemas inovadores relacionados à saúde pública. Um exemplo importante é a geração dinâmica de mapas de risco de doenças transmissíveis[1]. O mapeamento da distribuição geográfica de doenças transmissíveis é fundamental para programas de saúde pública. Entretanto, apenas um pequeno número de doenças transmissíveis de interesse está mapeado de forma completa em nível global. Como ilustração, Hay e colegas analisaram 355 doenças transmissíveis[21]. Destas, 174 tem uma justificativa forte pra o mapeamento. Das 174, apenas 7 foram mapeadas de forma completa. O uso de Big Data, além de facilitar o mapeamento através da integração de dados heterogêneos, permite o mapeamento de doenças transmissíveis em "tempo-real", através da integração de mapas estáticos de risco com relatórios de ocorrência atualizados continuamente[1].

BIG DATA EM GENÔMICA

Um genoma é o conjunto completo do DNA de um organismo, e contém quase todas as informações necessárias para construí-lo e mantê-lo vivo. Nos seres humanos, a totalidade da sequência do genoma é composto por mais de 3 bilhões de pares de bases (pb) de DNA, que são encontrados em todas as células de um organismo. O genoma normalmente é armazenado como um arquivo baseado em texto da ordem de 3GB (considerando um byte por pb).

O custo do sequenciamento do genoma diminuiu exponencialmente nos últimos anos[22], assim como o número de genomas seqüenciados e armazenados está a aumentar a um ritmo semelhante[23]. Até recentemente, sequenciar um genoma humano custava vários milhares de dólares norte-americanos. No entanto, a marca impressionante de US$1.000 para o sequenciamento de um genoma humano completo foi alcançado no início de 2014 pela empresa Illumina com a plataforma HiSeq X Ten[24]. O custo de sequenciamento deverá continuar caindo nos anos seguintes, o que cria amplas oportunidades para a investigação biomédica em geral e em particular para a medicina personalizada. Naturalmente, esta tendência irá produzir um forte aumento na quantidade de dados gerados.

Para ilustrar a natureza das reduções nos custos de sequenciamento de DNA ao longo dos anos, a figura I mostra uma reta hipotética (em escala logarítmica) que reflete a *Lei de Moore*. A lei de Moore é a observação de que, ao longo da história da computação, o número de transistores em circuitos integrados e, consequentemente, o desempenho dos computadores dobram aproximadamente a cada 2 anos. Melhorias tecnológicas que 'acompanham' a Lei de Moore são amplamente consideradas como de excelente desempenho, tornando-a útil como referencial de comparação.

Na Figura I podemos observar a evolução do custo médio do sequenciamento do genoma humano, de acordo com dados fornecidos pelo Instituto de Pesquisa do Genoma Humano dos EUA. Podemos observar que a redução do custo de sequenciamento ocorre de forma muito mais rápida do que a preconizada pela Lei de Moore, principalmente a partir de 2007. Cabe observar que, da mesma forma que o custo de geração de dados por sequenciamento de DNA decresce exponencialmente, o volume de dados gerados aumenta também de forma exponencial.

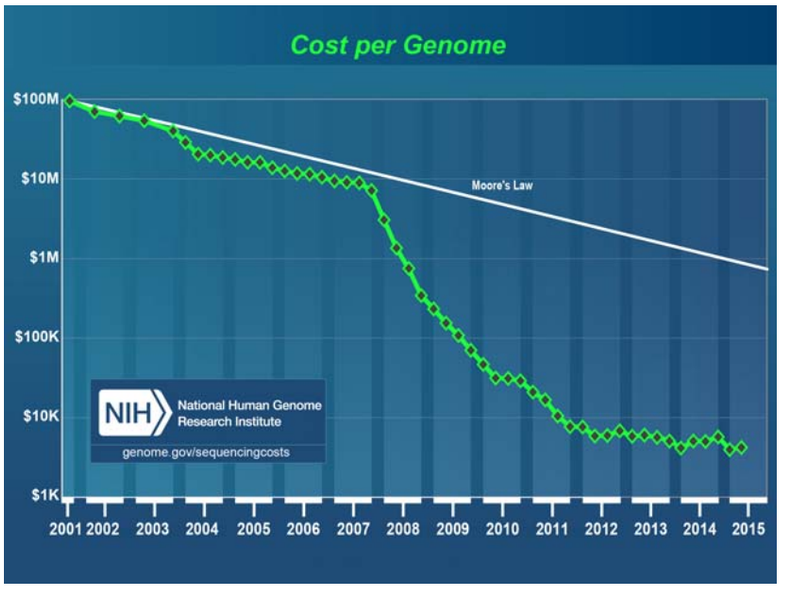

Figura I -Evolução do custo de sequenciamento do genoma. Fonte: US National Human Genome Research Institute[22]. Termos de uso disponíveis em http://www.genome.gov/sequencingcosts/

Desta forma, novas tecnologias de sequenciamento permitem que o sequenciamento completo de genomas em larga escala seja viável em termos de custo e tempo. A caracterização em larga escala do genoma humano envolve a geração e interpretação de um enorme volume de dados, em uma escala sem precedentes. O maior projeto deste tipo atualmente envolve o sequenciamento de 100.000 genomas, e é financiado pelo governo do Reino Unido[25]. Como exemplo de aplicação, é amplamente reconhecido que a geração de catálogos abrangentes das alterações somáticas em genomas do câncer, juntamente com o conhecimento detalhado da epigenética e de estados transcricionais destes genomas, vai permitir o desenvolvimento de novas estratégias eficazes de prevenção e de detecção precoce de diversos tipos de câncer. Além disso, torna possível o desenvolvimento de terapias apropriadas à subpopulação à qual o paciente pertence. A análise de um grande volume de dados genômicos do câncer também tem um grande potencial para determinar o prognóstico e orientar o tratamento da doença em seus estágios iniciais, baseado em evidências obtidas dos dados. O objetivo final desta caracterização em larga escala é medicina personalizada/de precisão para pacientes de câncer[7].

Vários projetos em andamento são voltados para a caracterização das mutações somáticas (substituições, inserções, deleções, etc.) em vários tipos de câncer, através da coleta de grandes volumes de dados, como o *The Cancer Genome Atlas (TCGA)*[26] e o *International Cancer Genome Consortium (ICGC)*[27]. Esses programas de integração de larga escala têm impulsionado não só a revolução tecnológica no campo da genômica, mas também estabeleceram normas comuns e de coordenação em escala internacional para assegurar que os dados gerados por diferentes grupos e países tenham o mesmo padrão de qualidade.

NUVENS COMPUTACIONAIS

Com o aumento da necessidade de armazenar dados e informação gerados por projetos de Big Data, diversas soluções computacionais têm sido propostas. Uma das mais promissoras é baseada no uso de nuvens computacionais[2-3, 8-10]. A computação em nuvem é o único modelo de armazenamento que pode fornecer a escala e a flexibilidade necessárias para o processamento de dados de sequenciamento de DNA[3], cuja taxa de avanço da tecnologia hoje em dia já excede a Lei de Moore, como discutido na seção anterior. Entretanto, embora as soluções de nuvem de diferentes organizações têm sido utilizadas em processamento de grandes volumes de dados de saúde, vários desafios permanecem, particularmente relacionados à segurança e privacidade dos dados pessoais, médicos e científicos.

A computação em nuvem pode ser definida como um modelo para permitir o acesso conveniente, sob demanda, a um *pool* compartilhado de recursos computacionais configuráveis (por exemplo, redes, servidores, armazenamento, aplicações e serviços) que podem ser rapidamente provisionados e liberados com um esforço de gerenciamento mínimo[28].

Entretanto, não basta ter acesso fácil a recursos computacionais através da nuvem. É preciso fazer uso eficiente (escalável) destes recursos, o que não é trivial. Um ambiente de computação paralela que é, ao mesmo tempo, simples de programar, eficiente, escalável e facilmente utilizável na nuvem computacional é conhecido como MapReduce[29]. MapReduce é um *framework* de computação paralela inventado pelo Google para o processamento de grandes conjuntos de dados. Os dados e os cálculos são distribuídos em um grande número de computadores, podendo processar grandes volumes de dados (e.g. da ordem de petabytes) por dia. Entretanto, a versão do MapReduce desenvolvida pela Google não está publicamente disponível. *Hadoop*[30] é a principal implementação de código aberto do MapReduce, e este sistema é usado amplamente por organizações governamentais, empresas e universidades.

Nuvens Computacionais e Genômica

Um exemplo de aplicação MapReduce para processar dados genômicos é o *Crossbow*[31]. Esta aplicação é um pipeline de software escalável para análise de genomas inteiros sobre *Hadoop*. Entretanto, é importante notar que atualmente já existem dezenas de aplicações MapReduce disponíveis para processar dados genômicos na nuvem computacional[10].

Cabe ressaltar que *Hadoop* não é o único ambiente de computação distribuído utilizado no processamento de dados genômicos. Uma alternativa promissora para *Hadoop* é o sistema *Spark*[32], que é um ambiente MapReduce voltado para a nuvem computacional. *Spark* oferece diversas extensões em relação ao modelo de computação do ambiente MapReduce original e é otimizado para processamento em memória, de alto desempenho. Cabe ressaltar que já existem pipelines descritos na literatura que utilizam *Spark*, em vez de *Hadoop*, para o processamento genômico[33].

É importante enfatizar que mesmo os laboratórios que não possuem expertise no desenvolvimento e adaptação de aplicações para nuvens computacionais podem se beneficiar deste ambiente computacional. Através do uso de portais especializados, disponíveis através da internet, usuários podem fazer uso de aplicações genômicas na nuvem de forma simples, sem necessidade de maiores conhecimentos técnicos sobre nuvens computacionais. Estes portais são baseados em plataformas específicas de código aberto, como o *Galaxy Cloudman*[34]. Cabe ressaltar que diversas organizações proveem acesso às nuvens computacionais através das plataformas supracitadas. Uma outra opção é a utilização de plataformas comerciais voltadas especificamente para o processamento de dados genômicos na nuvem computacional. Um exemplo é a plataforma *DNAnexus*[35], que é uma infraestrutura de nuvem voltada especificamente para o processamento de dados genômicos que faz uso de recursos computacionais da nuvem da Amazon.

Nuvens Computacionais e Saúde Pública

Através da análise de grandes volumes de dados e computação em nuvem, agências de saúde pública têm a oportunidade de observar ameaças emergentes à saúde pública em tempo real e proporcionar intervenções mais eficazes abordando as disparidades de saúde nas comunidades[36-38].

Um exemplo de estrutura de nuvem voltada para a saúde pública é o *Smarter Public Health Prevention System* (SPHPS)[36]. O SPHPS fornece relatórios em tempo real de potenciais ameaças à saúde pública para decisores, através da utilização de uma interface simples e eficiente, além de conectar pessoas com serviços necessários de saúde pessoal através do uso de plataformas móveis (celular, smatphone, tablet), visando promover e incentivar comportamentos saudáveis nas comunidades. O SPHPS implanta o conceito de Nuvem Virtual Privada (*Virtual Private Cloud*), que é definido como um espaço na nuvem dedicado a um único projeto ou entidade. Os recursos da nuvem são dedicados ao projeto, mas podem estar distribuídos em diversos locais geograficamente distribuídos e são acessíveis através de VPNs seguras[36].

DESAFIOS RELACIONADOS A BIG DATA E NUVENS COMPUTACIONAIS

Nesta seção, discutiremos alguns desafios relacionados a Big Data e nuvens computacionais, mantendo o foco em aplicações genômicas e de saúde pública. Os desafios que serão discutidos nesta seção são privacidade e segurança dos dados, transferência e armazenamento de dados, treinamento em ciência de dados e acesso a dados biomédicos.

Privacidade e segurança dos dados

Um dos desafios relacionados com nuvens públicas é o impedimento de se processar dados clínicos devido aos requisitos de privacidade e sigilo destes dados, com reflexos na legislação de diversos países (e.g. o capítulo de privacidade do *Health Insurance Portability and Accountability Act* - HIPAA, EUA). Por exemplo, um genoma contém informações altamente sensíveis que identificam um indivíduo. Quando a tecnologia avançar, eventualmente, ao ponto de fazer um sequenciamento completo de genoma se torne acessível para a população em geral, os indivíduos precisarão de garantias sobre o acesso à sua informação genômica. Com custo de sequenciamento de DNA caindo abaixo de US$ 1.000 por genoma, questões relacionadas à privacidade destes dados tornaram-se prementes[39]. Algumas plataformas recentes[40-41] implementam mecanismos para aumentar o nível de privacidade e segurança de acesso a dados clínicos na nuvem computacional, porém pesquisas adicionais sobre este tópico são necessárias.

Transferência e armazenamento de dados

Outro ponto crucial a ser atacado é a transferência de grandes volumes de dados, possivelmente através de uma rede de longa distância, para ser processado pela nuvem. Em muitos casos, o tempo necessário para transferir grandes volumes de dados para a nuvem computacional excede em muito o tempo necessário para processá-los[42]. Entretanto, alguns avanços recentes relacionados às questões de transferência de dados para nuvens são descritos na literatura. Exemplos são o Globus Online, que faz uso de um serviço otimizado de transferência de dados chamado GridFTP[43], protocolos otimizados para a transferência de dados em longas distâncias como o fasp[9], a sobreposição de comunicação e computação através do uso de *streaming*[42], e protocolos de comunicação par-a-par otimizados para a transmissão dados biológicos (e.g. Biotorrents[44]).

Outra possibilidade consiste no armazenamento prévio de grandes bases de dados na própria nuvem, minimizando a necessidade de transferências de dados. Por exemplo, a nuvem pública da Amazon disponibiliza um repositório com diversas bases públicas de dados genômicos que podem ser utilizadas por seus usuários de acordo com o modelo *Dados como Serviço (DaaS)*. É fundamental que nuvens computacionais voltadas para a

saúde pública e genômica sejam capazes de armazenar e disponibilizar estes dados adequadamente, de forma a minimizar a necessidade de transferência de grandes volumes de dados e a otimizar o desempenho das aplicações a serem executadas na nuvem.

Treinamento em Ciência de Dados

Um ponto fundamental para transformar em realidade a promessa do Big Data é o investimento em treinamento de cientistas e técnicos, em particular nas áreas relacionadas com o que se define atualmente como "Ciência de dados". Habilidades em demanda no âmbito da ciência de dados incluem ciência da computação, matemática e estatística, informática biomédica, biologia e medicina, entre outras, todas incorporadas do ponto de vista da ciência de dados. Cabe enfatizar a geração de grandes quantidades de dados, em conjunto com as complexas questões relacionadas que são colocadas, requerem equipes interdisciplinares para projetar os estudos e executar as análises de dados relacionadas[45], e este caráter interdisciplinar precisa ser levado em conta no planejamento de propostas de treinamento.

Acesso a dados biomédicos

Uma questão fundamental para o uso de Big Data em saúde pública é como habilitar a identificação, a acesso e as citações (i.e. o crédito) de dados biomédicos[45]. Inerente à descoberta de dados é a necessidade de um plano sustentável e escalável para criar e manter um sistema de descoberta que permite que aos pesquisadores encontrar facilmente e citar dados biomédicos. De fato, a sustentabilidade e a escalabilidade são duas questões interligadas que devem ser abordadas para que os avanços possibilitados pelo uso de Big Data em Saúde tenham um efeito duradouro. Um proposta neste sentido foi feita pela iniciativa *Big Data to Knowledge* (BD2K), do NIH/EUA, de criar um Índice de descoberta de dados (*Data Discovery Index* - DDI)[45]. O DDI deverá implantar abordagens avançadas para a busca, integração e visualização dos dados.

Entre os desafios relacionados à acessibilidade dos dados está uma melhor forma de expandir a disponibilidade e uso de prontuários eletrônicos de saúde[45]. Dados clínicos de prontuários eletrônicos, juntamente com dados de saúde individuais capturados por vários dispositivos pessoais, oferecem consideráveis oportunidades para o avanço da pesquisa clínica e biomédica. No entanto, ao contrário de outros tipos de dados usados em pesquisas biomédicas, dados clínicos são normalmente capturados fora dos ambientes tradicionais de pesquisa e devem então ser readaptados para tal uso. Esta readequação levanta questões importantes relacionadas à integração de dados heterogêneos e à proteção da privacidade do paciente.

Uma abordagem possível para este desafio é a utilização de biobancos, que são plataformas voltadas para o armazenamento e acesso escalável de material biológico de indivíduos para utilização clínica e em pesquisa[40]. No entanto, sistemas de software para biobancos tradicionalmente gerem apenas metadados associados a amostras, tais como pseudo-identificadores para os pacientes, informações de coleta de amostras, ou informações relacionadas a estudos específicos. Normalmente, tais sistemas não conseguem lidar com o requisito para, ao lado de metadados, também armazenar e analisar dados de diversos tipos como, por exemplo, os dados genômicos, os quais podem variar desde poucos Megabytes (por exemplo, informações de vetores SNP) a centenas de Gigabytes por amostra (sequenciamento do genoma completo com alta cobertura). Entretanto, a integração com nuvens computacionais está mudando este cenário[40].

CONCLUSÃO

O enfrentamento os desafios associados com Big Data em saúde pública e genômica grande deve necessariamente envolver todas as áreas relacionadas com o tema ciência de dados. Embora estes desafios sejam complexos, eles se mostram de resolução possível considerando as diversas tecnologias de Big Data disponibilizadas recentemente e mencionadas neste artigo. Os desafios devem ser abordados necessariamente com uma visão interdisciplinar e interação próxima de pesquisadores das disciplinas relacionadas às ciências de dados, saúde pública e genômica.

Como fica claro através da análise de diversos trabalhos citados neste artigo de revisão, Big Data em saúde pública e em genômica é uma realidade hoje. Este fato reforça a importância das tecnologias voltadas para o processamento e análise de Big Data citados neste trabalho. Além disso, também são citados neste manuscrito diversos trabalhos que mostram que a nuvem computacional é uma plataforma adequada para o processamento de grandes volumes de dados, e pode ser usada também em diversas aplicações relacionadas à saúde pública e à genômica.

Ao descobrir novas associações e pela compreensão dos padrões e tendências nos dados, a análise de Big Data tem o potencial para melhorar o atendimento prestado à população, salvar vidas e reduzir custos. Acreditamos que o processamento eficaz de Big Data, através da utilização das nuvens computacionais, tem o potencial de aumentar de forma significativa o alcance e a eficácia de diversas ações de Saúde Pública no Brasil.

REFERÊNCIAS

1. Hay SI, George DB, Moyes CL, Brownstein JS. Big data opportunities for global infectious disease surveillance. PLoS Med. 2013;10(4):e1001413.
2. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Inf Sci Syst. 2014;*2*(1):3.
3. Costa FF. Big data in biomedicine. Drug discovery today. 2014; 19(4):433-40.
4. Hey AJ, Trefethen AE. The data deluge: an e science perspective. In: Grid computing: making the global infrastructure a reality. 2003. p.809-24.
5. Genbank Statistics. [Citado em 2015 Set 8]. Disponível em http://www.ncbi.nlm.nih.gov/genbank/statistics
6. Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, et al. Detecting novel associations

in large data sets. Science. 2011; 334(6062):1518-24.
7. Chin L, Andersen JN, Futreal PA. Cancer genomics: from discovery science to personalized medicine. Nat Med. 2011;17(3):297-303.
8. Dai L, Gao X, Guo Y, Xiao J, Zhang Z. Bioinformatics clouds for big data manipulation. Biol Direct. 2012;7(1):43.
9. Marx V. Biology: the big challenges of big data. Nature. 2013;498(7453): 255-60.
10. O'Driscoll A, Daugelait J, Sleator RD. 'Big data', Hadoop and cloud computing in genomics. J Biomed inform. 2013;46(5):774-81.
11. Salathé M, Bengtsson L, Bodnar TJ, Brewer DD, Brownstein JS, Buckee, et al. Digital epidemiology. PLoS Comput Biol. 2012;8(7):e1002616.
12. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data**.** Nature. 2009;457(7232):1012-4.
13. Chan EH, Sahai V, Conrad C, Brownstein JS. Using web search query data to monitor dengue epidemics: a new model for neglected tropical disease surveillance**.** PLoS Negl Trop Dis. 2011;5(5):e1206.
14. Signorini A, Segre AM, Polgreen PM. The use of twitter to track levels of disease activity and public concern in the U.S during the influenza a H1N1 pandemic. PLoS One. 2011;6(5):e19467.
15. Butler D. When google got flu wrong. Nature. 2013;494(7436):155.
16. White RW, Tatonetti NP, Shah NH, Altman RB, Horvitz E. Web-scale pharmacovigilance: listening to signals from the crowd. J Am Med Inform Assoc. 2013; 20(3):404-8.
17. Harpaz R, Vilar S, Dumouchel W, Salmasian H, Haerian K, Shah NH, et al. Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions**.** J Am Med Inform Assoc. 2013;20(3):413-9.
18. Bian J, Topaloglu U, Yu F. Towards large-scale twitter mining for drug-related adverse events. In: Proceedings of the 2012 ACM international workshop on Smart health and wellbeing; 2012 Oct 29-Nov 2; Maui, HI, USA. p.25-32.
19. Freifeld CC, Brownstein JS, Menone CM, Bao W, Filice R, Kass-Hout T, Dasgupta N. Digital drug safety surveillance: monitoring pharmaceutical products in twitter. Drug Saf. 2014;37(5):343-50.
20. Sampathkumar H, Chen XW, Luo B. Mining adverse drug reactions from online healthcare forums using hidden Markov model. BMC Med Inform Decis Making. 2014;14(1):91.
21. Hay SI, Battle KB, Pigott DM, Smith DL, Moyes CL, Bhatt S, et al. Global mapping of infectious disease. Philos Trans R Soc Lond B Biol Sci. 2013;368(1614):20120250.
22. Wetterstrand KA. DNA Sequencing costs: data from the NHGRI genome sequencing program (GSP) [cited 2015 Set 8]. Available from: http://www.genome.gov/sequencingcosts/
23. Kahn SD. On the future of genomic data. Science. 2011;331(6018):728-9.
24. Illumina Inc. Illumina introduces the HiSeq X Ten sequencing system. [cited 2015 Set 9]. Available from: http://investor.illumina.com/phoenix.zht ml?c=121127&p=irol-newsArticle
25. England G. The 100.000 Genomes Project. [cited 2015 Set 8]. Available from: http://www.genomicsengland.co.uk/
26. TCGA Project. [cited 2015 Set 9]. Available from: http://cancergenome.nih.gov/.
27. ICGC Consortium. [cited 2015 Set 9]. Available from: http://icgc.org/
28. Mell P, Tim G. The NIST definition of cloud computing. Communic ACM. 2010;53(6):50.
29. Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. Communica ACM. 2008; 51(1):107-13.
30. White T. Hadoop: the definitive guide. O'Reilly; 2012.
31. Langmead B, Schatz MC, Lin J, Pop M, Salzberg SL. Searching for SNPs with cloud computing. Genome Biol. 2009;10(11):R134.
32. Zaharia M, Chowdhury M, Franklin MJ, Shenker S, Stoica I. Spark: cluster computing with working sets. In: Proceedings of the 2nd USENIX conference on Hot topics in cloud computing; 2010 Jun 22-25; Boston, MA.
33. Massie M, Nothaft F, Hartl C, Kozanitis C, Schumacher A, Joseph AD, et al. ADAM: genomics formats and processing patterns for cloud scale computing. Technical Report UCB/EECS-2013-207, EECS Department, University of California, Berkeley; 2013.
34. Afgan E, Baker D, Coraor N, Chapman B, Nekrutenko A, Taylor J. Galaxy CloudMan: delivering cloud compute clusters. BMC bioinformatics. 2010;11(Suppl 12):S4
35. DNAnexus Inc. The DNAnexus platform. [cited 2015 Set 8]. Available from: http://www.dnanexus.com/
36. Jalali A, Olusegun AO, Bell CM. Leveraging cloud computing to address public health disparities: an analysis of the SPHPS. Online J Pub Health inform. 2012;4(3).
37. Piette JD, Mendoza-Avelares MO, Ganser M, Mohamed M, Marinec N, Krishnan S. A preliminary study of a cloud-computing model for chronic illness self-care support in an underdeveloped country. Am J Prev Med. 2011;40(6):629-32.
38. Kuo AM. Opportunities and challenges of cloud computing to improve health care services. J Med Internet Res. 2011;13(3):e67.
39. Ayday E, De Cristofaro E, Hubaux JP, Tsudik G (2015). Whole Genome Sequencing: Revolutionary Medicine or Privacy Nightmare? IEEE Computer. 2015;(2):58-66.
40. Bessani A, Brandt J, Bux M, Cogo V, Dimitrova L, Dowling J, et al. BiobankCloud: a platform for the secure storage, sharing, and processing of large biomedical data sets. Proceedings of the 1st Int. In: Workshop on Data Management and Analytics for Medicine and Healthcare (DMAH 2015); 2015 Sept 4; Kohala Coast. Hawaii, US.
41. DNAnexus Inc. Compliance with HIPAA, CLIA, dbGaP, EU Privacy, and ISO 27001 on DNAnexus. [cited 2015 Set 9]. Available from: https://www.dnanexus.com/papers/Compliance_White_Paper.pdf
42. Issa SA, Kienzler R, El-Kalioby M, Tonellato PJ, Wall D, Bruggmann R, Abouelhoda M. Streaming support for data intensive cloud-based sequence analysis. Biomed Res Int. 2013;2013:791051.
43. Foster I. Globus online: Accelerating and democratizing science through cloud-based services. IEEE Internet Computing. 2011:15(3):70-3.
44. Langille MG, Eisen JA. BioTorrents: a file sharing service for scientific data. PLoS One. 2010;5(4):e10071.
45. Margolis R, Derr L, Dunn M, Huerta M, Larkin J, Sheehan J, et al. The National Institutes of health's big data to knowledge (BD2K) initiative: capitalizing on biomedical big data. J Am Med Inform Assoc. 2014;21(6):957-8.