FIAP ON
ENSINO A DISTÂNCIA

BIG DATA & ANALYTICS

GOVERNANÇA DE DADOS - BIG DATA

Anderson Paulucci e Leandro Rubim

CAPÍTULO 6

LISTA DE FIGURAS

SUMÁRIO

6.1 GOVERNANÇA DE DADOS (BIG DATA)

A única maneira sustentável para transformar dados em *insights* de negócios é com um processo de governança de dados bem estruturado, e ao longo deste capítulo vamos entender que “bem estruturado” não significa um rígido controle de governança.

Dessa forma, entramos aqui em um grande dilema ao abordar um assunto tão importante para a evolução de Big Data. Afinal, se temos possibilidade de processar e analisar os dados não estruturados e aceitar a “confusão” dos dados, afrouxando a qualidade, comparando com os processos burocráticos da modelagem relacional, em prol de armazenar mais dados, como fica a questão da governança?

As mudanças nunca mais vão acontecer de forma tão lenta como hoje! A explosão de dados com Big Data, que é apenas um aperitivo para o que temos pela frente com *IoT* e a onda de *Deep Learning*, é um grande desafio para a governança de dados. Estamos ainda em um momento de transição no processo de disrupção para o Digital, pode parecer algo novo, mas já vimos isso com as aplicações empresariais no fim dos anos 1990 (guardadas, claro, as devidas proporções), ignore a concordância verbal “vimos”, caso você seja um nativo da era Digital. Muitas empresas passaram por grandes desafios com múltiplos ERPs, CRMs e sistemas corporativos, algumas empresas estavam estruturando processos de consolidação do mercado, com fusões e aquisições e elaborando silos de informações devido à complexidade das integrações, criando um verdadeiro colapso para a arquitetura de dados das empresas. Em muitos aspectos, esse fenômeno (pelo menos parcialmente) levou à necessidade de qualidade dos dados, governança de dados e administração de dados no mundo da TI. Foi uma época de enorme crescimento de dados, e tivemos que descobrir uma maneira de gerenciar a informação, priorizando assuntos de governança.

Como mencionado no capítulo “Data Lake – Uma nova abordagem para DW”, é um grande desafio manter as “águas” do lago claras e passíveis de navegação, e, para isso, a adoção de processos de governança aplicados ao ciclo de vida dos dados é fundamental. Por outro lado, também é preciso garantir que esses processos não impactem a agilidade de analisar e processar Big Data. Um

grande esforço de compreensão sobre as atuais atribuições dos guardiões dos dados, conhecidos como Data Stewards ou administradores de dados, é importante.

Uma política base para governança de dados deve compreender os seguintes pontos:

- Identificação de perfil e descoberta de dados para encontrar problemas na qualidade de dados.
- Monitoramento da qualidade e da linhagem de dados para administrar possíveis problemas de qualidade de dados por toda a empresa e garantir que as expectativas de qualidade de dados sejam atendidas.
- Gerenciamento de dados mestres para estabelecer uma visão única e fidedigna de clientes, produtos ou qualquer outro domínio de dados.
- Um glossário de negócios para definir padrões de termos de negócios e assegurar as comunicações claras sobre a integração dos dados.
- O gerenciamento da vida útil das informações de aplicativos para gerenciar o crescimento dos dados, desativar com segurança os sistemas e aplicativos legados.

“Governança de Dados é o exercício de autoridade e controle (planejamento, monitoramento e execução) sobre o gerenciamento de ativos de dados. A função de governança de dados guia como todas as outras funções da gestão de dados são realizadas. Governança de dados é de alto nível, ou seja, é gestão estratégica de dados na esfera executiva.” DAMA

Através do atual guia DAMA-DMBOK® podemos compreender que a Gestão de Dados é uma disciplina formada pelo conjunto de dez funções de gerenciamento de dados integradas. A integração dessas funções é feita pela Governança de Dados, por essa razão, ela está localizada como elemento central do *framework* do DAMA-DMBOK®.

Figura 6.1 – Governança de dados
Fonte: <http://dama.org/content/dama-dmbok>.

Abaixo, o conjunto das funções definidas pelo guia DAMA-DMBOK® para Gestão de Dados:

Gestão da Arquitetura de Dados	Função responsável por definir as necessidades de dados e alinhar os mesmos com a estratégia de negócio da empresa.
Gestão do Desenvolvimento de Dados	Função responsável pelas atividades de modelagem e implementação das estruturas dos dados dentro do ciclo de vida do desenvolvimento dos sistemas de informação.
Gestão de Operações de Dados	Função responsável por manter armazenados os dados ao longo do seu ciclo de

	vida.
Gestão da Segurança dos Dados	Função responsável por definir e manter as políticas de segurança da informação da empresa.
Gestão de Dados Mestres e Dados de Referência	Função responsável por definir e controlar atividades para garantir a consistência e disponibilização de visões únicas dos principais dados reutilizados na empresa.
Gestão de *Data Warehousing* e *Business Intelligence*	Função responsável por definir e controlar processos para prover dados de suporte à decisão, geralmente disponibilizados em aplicações analíticas.
Gestão da Documentação e Conteúdo	Função dedicada a planejar, implementar e controlar atividades para armazenar, proteger e acessar os dados não estruturados das empresas.
Gestão de Metadados	Os metadados representam o significado dos dados. Esses significados correspondem tanto ao conteúdo técnico do dado, obtido através das informações sobre estrutura, formato, tamanho e restrições, como a informações sobre definições e conceitos.
Gestão da Qualidade dos Dados	Função responsável por promover, medir, avaliar, melhorar e garantir a qualidade dos dados da empresa.

Governança de dados vai além da gestão de dados operacionais e pode ser construída com base em quatro pilares:

Estratégia	Contempla o alinhamento dos objetivos estratégicos e de negócio da organização com o

	conjunto de processos, dados e tecnologias relacionados ao uso e consumo da informação, de forma que sua utilização traga vantagens competitivas à empresa. Suas principais atividades são: ✓ Entendimento dos objetivos de negócio da empresa. ✓ Identificação das necessidades futuras. ✓ Elaboração de um plano estratégico para atendimento dos requisitos futuros do negócio. ✓ Reavaliação periódica do planejamento.
Qualidade	Planejamento, implementação e acompanhamento de processos que garantam o atendimento das necessidades dos consumidores de informação da organização. O conceito da qualidade está intimamente ligado à expectativa do consumidor em relação ao produto informação, que é o que direciona a definição do nível de qualidade adequado, considerando o grau de atendimento das suas reais necessidades, além do custo de captação e manutenção da informação. Aplicando o mesmo conceito para garantir a qualidade da informação, vários de seus aspectos – ou dimensões – devem ser considerados: a qualidade vai além da integridade, ou exatidão. É preciso também avaliar, por exemplo, a atualidade, a facilidade de uso, a segurança de acesso, entre outras características. É no processo de avaliação dessas dimensões que entram as métricas: regras de avaliação da qualidade para cada dimensão, sejam de aferição subjetiva ou objetiva. Uma vez

	definidas as métricas, elas passam a ser aplicadas regularmente, constituindo-se um termômetro imprescindível para monitorar a qualidade dos dados da empresa e o resultado das ações de melhoria adotadas ao longo do tempo. As principais atividades dessa área são: ✓ Entendimento das necessidades de Qualidade de Dados dos consumidores. ✓ Manutenção do nível de Qualidade de Dados. ✓ Definição e revisão das métricas de Qualidade de Dados. ✓ Monitoramento de métricas de Qualidade de Dados. ✓ Ações para melhoria do nível de qualidade das informações. ✓ Data Stewardship. ✓ Oferta de serviços padronizados de Qualidade de Dados. ✓ Disseminação dos conceitos de Qualidade de Dados.
Gestão	Estudo, negociação e acompanhamento da adoção de políticas e melhores práticas de produção e consumo da informação no ambiente corporativo. Suas principais atividades são: ✓ Definição de políticas e padrões para coleta, armazenamento e utilização das informações. ✓ Definição de procedimentos para acesso e proteção dos dados. ✓ Atendimento das questões regulatórias.

Arquitetura	Gestão do macroambiente de captação e manutenção das informações na organização. Suas principais atividades são: ✓ Mapeamento do ciclo de produção da informação. ✓ Integração de dados. ✓ Atendimento dos requisitos de negócio. ✓ Padronização e gerenciamento de metadados. ✓ Atendimento dos requisitos de tecnologia. ✓ Padronização e gerenciamento da modelagem de dados.

O principal objetivo da governança de dados é prover a gestão do ativo dados e informação.

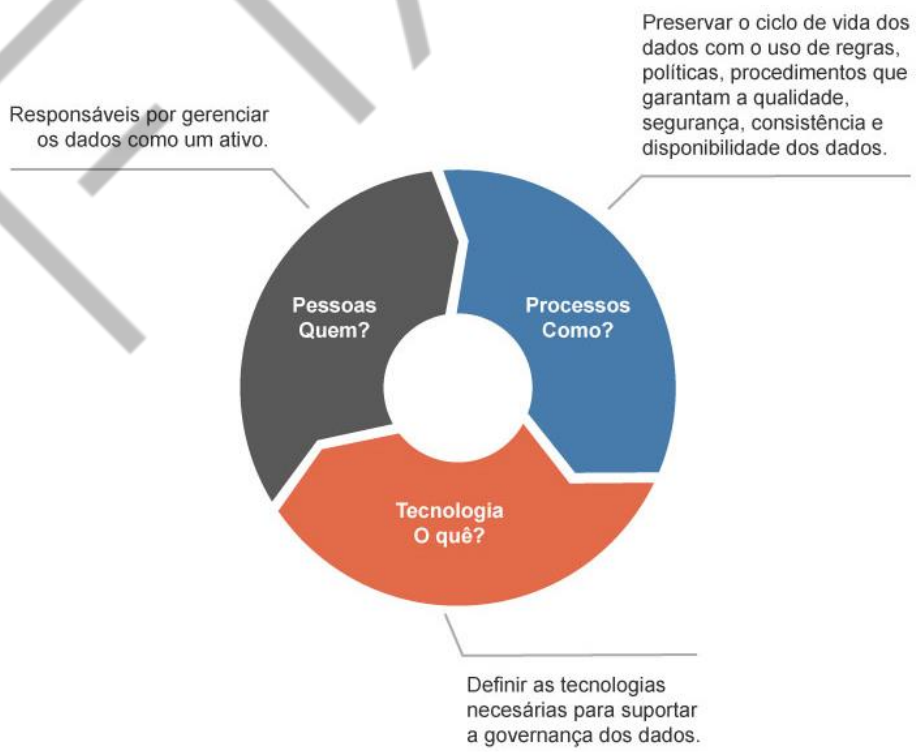

Figura 6.2 – Responsabilidade de Negócios
Fonte: Elaborado pelo autor (2017).

Considerando o ciclo de vida do dado e da informação tal como expressado no DAMA-DMBoK®, é possível identificar claramente as fases de responsabilidade de TI e de Negócios. De forma mais objetiva, podemos definir que Planejar, Especificar e Disponibilizar a estrutura para receber o dado são fases que estão na esfera de coordenação da TI. Já Criar ou Adquirir, Manter e Usar, Arquivar e Recuperar e, por fim, Eliminar são fases inerentes ao negócio. O ciclo de vida do dado e da informação nos permite identificar a responsabilidade de cada um.

6.1.1 Data Stewards ou Administrador de Dados

Um administrador de dados é responsável pela definição de políticas de uso, segurança e manutenção da qualidade dos dados, exercendo funções de gestão dos dados com todos os seus elementos e metadados, conforme determinado através de iniciativas de governança de dados estabelecida, atuando como uma ligação entre o departamento de TI e a área de negócio de uma organização.

Figura 6.3 - Analogia a segurança e manutenção da qualidade dos dados
Fonte: Banco de imagens Shutterstock

Gartner diz que as organizações devem estabelecer regras de administração de dados para melhorar a qualidade dos dados. Segundo Gartner, empresas europeias classificaram a qualidade de dados deficiente como o segundo maior problema de inteligência de negócios.

As organizações que se esforçam para melhorar a qualidade dos dados devem considerar a nomeação de administradores de dados, afirma Gartner. O sucesso da administração de dados exige que as organizações se movam em direção a uma cultura que considera os dados como um ativo competitivo ao invés de um mal necessário e que defina metas claras para a melhoria da qualidade dos dados.

“A qualidade dos dados é uma questão de negócios, não uma questão de TI, e exige do negócio assumir a responsabilidade e impulsionar melhorias”, disse Ted Friedman, vice-presidente de pesquisa do Gartner. A nomeação de administradores de qualidade de dados ajuda as organizações a alcançar os objetivos de melhoria da qualidade dos dados. Esses indivíduos devem ser considerados peritos em assuntos para seus departamentos e agir como *trustees* de dados, em vez de proprietários deles. Eles irão garantir que a qualidade será mantida para suportar processos de negócios.

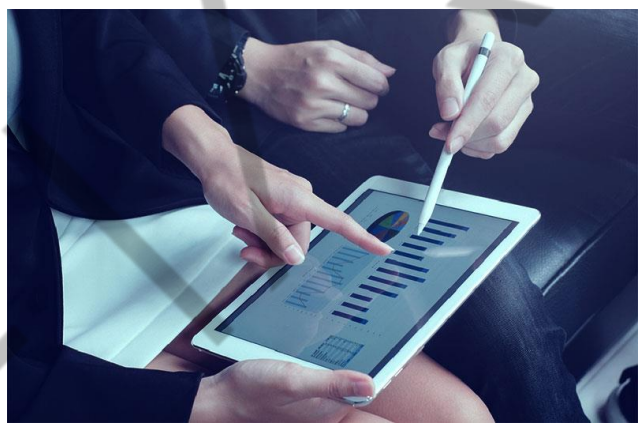

Figura 6.4 - Analogia a administradores de qualidade de dados
Fonte: Banco de imagens Shutterstock

Por exemplo, um especialista em marketing do departamento de marketing da empresa poderia atuar como administrador de dados no programa de melhoria da qualidade dos dados, mantendo os dados de marketing completos, corretos, consistentes, confiáveis e não redundantes. Nessa função, eles têm a responsabilidade de assegurar as informações relevantes para o marketing seguir os padrões de qualidade dos dados corporativos.

O desafio que já era enorme para a governança de dados ficou ainda maior com Big Data e os administradores de dados, apresentar novas soluções para estas complicações não será tão simples como as transações de dados estruturados.

Isso se deve à natureza altamente variável das estruturas de Big Data, que podem incluir uma combinação de tipos de dados estruturados e não estruturados, dados de transação, arquivos de logs dos sistemas e da rede, informações de sensores, registros de pesquisa e métricas na web e texto de redes sociais, apenas para citarmos alguns exemplos. Esses dados são provenientes de sistemas externos, acrescentando outro fator complicador para os administradores de dados, que não podem exercer nenhum controle sobre a qualidade e a consistência da informação à medida que ela está sendo criada.

Os cientistas de dados podem não concordar com a limpeza, os ajustes ou a consolidação dos dados, porque os esforços de limpeza podem distorcer os resultados dos algoritmos avançados de análise que eles estão desenvolvendo. Os cientistas dizem que não é um lugar para a administração de dados, mas sim para a descoberta.

Atenção, não se engane com a posição dos cientistas de dados em relação à necessidade de governança dos dados, eles possuem competências que os diferenciam de outros profissionais, usam suas habilidades para explorar *advanced analytics* e podem “tirar leite de pedra”; este argumento não é válido para os dados on-line de um processo de negócio, por exemplo.

Considere um cadastro de clientes em que o atributo idade não esteja devidamente preenchido, e uma análise de negócio dependa dessa informação, para um analista da área de negócio não há muito o que fazer e a alternativa de enriquecer o cadastro a partir da interação com o cliente não seria efetiva, a menos que se considerasse ações de médio e longo prazos. Porém, para a ciência de dados é possível adicionar outras métricas no modelo analítico, como, por exemplo: Quantos filhos tem o cliente? ou Qual é o ano de sua formação acadêmica? ou Há quantos anos o cliente está casado? ou Qual é o primeiro número do RG?... Enfim, essas perguntas podem ajudar a compor um algoritmo que consiga fazer uma inferência da faixa etária do cliente (30-40 anos) e talvez isso baste para o negócio.

Ou considere outro exemplo baseado em internet das coisas (IoT), em que aparentemente os sensores estão apresentando dados imprecisos ou nulos, à medida que os dados são processados em um cenário de Big Data, com centenas de milhares de sensores e eventos sendo analisados, é possível rapidamente entender alguns padrões e, através da estatística, identificar os *outliers* (valores

atípicos) e compreender a inconsistência sem a necessidade de controles mais rígidos na coleta dos dados. Dessa maneira, os dados podem ser carregados mesmo com inconsistência e quanto maior a granularidade dos dados e número de sensores coletando dados, melhor; a prioridade deve ser coletar os dados, a tendência de montar um modelo mais preciso não depende de validações rígidas no fluxo de armazenamento, como vimos, pode ser tratada com estatística em tempo de análise.

Mas nem sempre este será o melhor caminho, afinal, os cientistas de dados dependem de dados confiáveis para evoluir as atividades; portanto, o conhecimento desses profissionais não resolve todos os problemas de qualidade dos dados, e quando estamos tratando de aplicações on-line que não passam por análise de cientistas de dados e que devem manter uma segurança no fluxo de integração contínua, a governança deve ser sempre tratada com grande atenção.

Normalmente os administradores de dados vêm de áreas funcionais da empresa, e eles trabalham em TI para garantir a qualidade de Big Data e as políticas para armazenar e acessar os dados. Eles também têm as habilidades pessoais e políticas necessárias para obter consenso organizacional sobre quais dados devem ser “limpos” e armazenados, quem pode acessá-los e quando devem ser expurgados.

Big Data propõe uma abordagem com o uso da persistência poliglota, abrindo caminho para modelos Not Only SQL ou simplesmente NoSQL (Key/Value, Column Family, Graph, Documents), os atuais administradores de dados que estão habituados a tratar a modelagem Entidade e Relacionamento com *schemas* bem definidos, se deparam com desafios para entender o modelo que será em sua maior parte definido dinamicamente pelo desenvolver, “sem a necessidade” de uma alteração burocrática do modelo lógico e físico, fora da aplicação, com o uso do conceito *schema-less.*

Aprender outras estruturas de modelagem é fundamental para os Data Stewards tradicionais, a governança de dados nunca será a mesma com a redundância dos dados prevista na modelagem NoSQL, afinal ela propõe que o aumento do volume de dados resultante da desnormalização dos dados não é um problema, o desafio é resolver os JOINs na execução, o modelo orientado a consulta será uma grande mudança de *mindset* para os administradores de dados. Os

desenvolvedores não podem ter total autonomia para criar e alterar os modelos, afinal, eles não têm a preocupação-chave com a governança, e este deve ser sempre o papel do administrador de dados: acompanhar os ciclos de evolução do código da aplicação para apoiar os engenheiros de dados na definição de qual seria o melhor modelo de persistência.

6.1.2 *Lean Manufacturing* – Uma visão de governança de dados para a era da agilidade

Quando ouvimos expressões como “processos e procedimentos”, imediatamente pensamos em “burocracia”, pensamos nisso como os inimigos da agilidade.

Figura 6.5 - Analogia a burocracia
Fonte: Banco de imagens Shutterstock

As empresas estão suplicando agilidade para a TI, que é um grande “gargalo” para as organizações tradicionais, e, como vimos, os esforços de governança dos dados invariavelmente serão tratados pela TI com a ajuda das áreas de negócio.

Proponho aqui um breve entendimento sobre uma metodologia de produção criada pela Toyota que se tornou uma filosofia de gestão amplamente adotada no mundo todo. *Lean Manufacturing* ou Manufatura Enxuta foi desenvolvida por Taiichi Ohno, executivo da Toyota durante o período de reconstrução do Japão após a Segunda Guerra Mundial. O termo foi popularizado por James P. Womack e Daniel

T. Jones no livro *A Mentalidade Enxuta nas Empresas Lean Thinking: Elimine o Desperdício e Crie Riqueza*.

A partir deste, vários conceitos atuais foram criados, como, por exemplo, Lean Startup e Design Thinking.

A abordagem de governança de dados tradicionais requer grande esforço e muito trabalho, mas pode não representar resultados satisfatórios a longo prazo. A governança ágil/enxuta, por outro lado, está focada em capacitar as pessoas e motivá-las a fazer as coisas certas.

O especialista Scott Ambler, do Instituto Agile Data, aponta alguns passos importantes para um modelo ágil de governança de dados. Destaco abaixo os cinco principais:

	Ativos Corporativos Valorizados	As orientações (como convenções de design de banco de dados, diretrizes de estilo de modelagem, convenções de nomenclatura de dados e diretrizes de design de relatório), definições de metadados e ativos reutilizáveis, como estruturas e componentes, serão adotadas caso se considere que agregam valor aos desenvolvedores. Devemos torná-las tão fáceis quanto possível para que os desenvolvedores as utilizem e, o mais importante, aproveitem sua infraestrutura de TI corporativa. Quando os padrões de dados são sensíveis, fáceis de entender e de fácil acesso, há uma chance significativamente maior de que as pessoas realmente seguirão os padrões na prática. Se forçarmos as pessoas a se conformarem com os padrões burocráticos, isso pode ser muito oneroso para o processo, então reduz a chance de que eles realmente o façam .
	Incluir profissionais de dados como	Quando o grupo DM (Data Management) é externo às equipes de projeto, pode promover uma mentalidade de “eles contra nós” dentro de uma

	participantes ativos em equipes de desenvolvimento	organização de TI. Não é necessário ter um grupo externo para executar atividades de governança de dados, em vez disso, profissionais de dados individuais podem atuar com responsabilidades de gestão de dados em equipes de desenvolvimento de forma colaborativa e oportuna. Este é um dos conceitos fundamentais do método Agile Data.
	Educar os desenvolvedores	Os desenvolvedores precisam entender por que seus esforços de MDM (Master Data Management) são importantes, quais são os benefícios e como trabalhar em conjunto com sua equipe de DM. Quando eles sabem por que algo precisa ser feito, e como fazê-lo de forma eficaz, as chances de que eles realmente façam isso são muito maiores.
	Pipeline do Projeto guiado por Negócios	Devemos nos preocupar em que as atividades de TI com a gestão de dados estejam alinhadas com os objetivos de negócio. Infelizmente, diversas estratégias tradicionais de governança de dados muitas vezes parecem refletir as prioridades dos burocratas de dados, não as prioridades do negócio, resultando em repositórios de dados, com diversas barreiras e dados subutilizados.
	Conformidade integrada	É melhor construir a conformidade em seus processos do dia a dia, ao invés de ter um processo de conformidade separado que muitas vezes resulta em sobrecarga desnecessária. A automação é importante. Por exemplo, ao invés de realizar revisões para garantir que as equipes de desenvolvimento sigam as convenções de dados corporativos, um esforço demorado e caro, por que

		não executar uma ferramenta de análise de código estático sobre os esquemas de banco de dados em uma base regular, para garantir que as convenções de nomenclatura de dados sejam seguidas?

Big Data começou como uma solução de arquitetura para aumentar volumes e variedade dos dados, e está se tornando um problema econômico para as organizações que já têm mais dados do que podem gerenciar e estão lutando com o custo e a complexidade de manter escalabilidade para o gerenciamento. Governança ágil é saber quando e como gastar dinheiro em dados. O que acontece se você limpar e cozinhar um peixe e após a primeira mordida perceber que ele tem um gosto ruim? Big Data são cargas de caminhões de peixes sendo despejados no Data Lake todos os dias, como encontrar rapidamente os melhores peixes para degustação?

A IBM publicou um *paper* intitulado *"The IBM Agile Information Governance Process* (2014) " para abordar a necessidade de uma metodologia ágil para a governança de dados. Propõe uma abordagem alternativa sobre como entregar as iniciativas de Big Data com projetos-piloto, mas ainda ser capaz de abordar os principais problemas de governança.

Figura 6.6 – IBM Agile Information Governance Process
Fonte: <https://www.flickr.com/photos/29986804@N08/10396735315>,

O IBM Agile Information Governance Process consiste em seis etapas e três fases distintas. Na fase Plano (*Plan*), as equipes de governança de dados definem o problema do negócio, obtêm patrocínio executivo, alinham com as equipes e entendem o risco e o valor dos dados. Na fase Ato (*Act*), as organizações implementam um ou mais projetos baseados em casos de uso comum. Finalmente, na fase Avalia (*Assess*), as equipes de governança de dados medem os resultados.

Figura 6.7 - Analogia a loop contínuo
Fonte: Banco de imagens Shutterstock

O IBM Agile Information Governance Process é construído como um *loop* contínuo. À medida que as equipes de governança de dados medem resultados em um projeto, elas começam de novo definindo o problema de negócios que pode gerar projetos adicionais.

Uma notável diferença entre a governança da velha escola e a governança ágil é a suposição de que os projetos irão entrar e começar a usar os dados sem necessariamente se preocupar com qualidade e consistência. Como afirma o *white paper*:

- Os dados podem ser carregados como estão porque os elementos de dados críticos e as relações podem não ser totalmente compreendidos.
- Os dados podem ser semiestruturados ou não estruturados e sujeitos a uma exploração adicional, portanto, os elementos de dados críticos podem mudar iterativamente.

- Os Stewards podem gerenciar apenas uma porcentagem menor de dados, devido a altos volumes e/ou velocidade.

Na “Economia baseada em Dados”, eventualmente temos que tomar atalhos. Governança de dados ágil não é apenas sobre os atalhos ou chegar a uma resposta mais rapidamente, trata-se do padrão de governança certo para cada situação. Às vezes, as técnicas de Big Data podem ser usadas para resolver problemas de governança.

REFERÊNCIAS

Agile Data. **Agile Data Techniques**. Disponível em: <http://agiledata.org>. Acesso em: nov. 2016.

DAMA International Guide to Data Management Body of Knowledge (DAMA DMBOK®). **Body of Knoledge.** Disponível em: <http://dama.org>. Acesso em: nov. 2016

IBM Software. **"The IBM Agile Information Governance Process"**. Maio de 2014. Disponível em: http://www.cio.co.uk/cmsdata/whitepapers/3534176/governanace.pdf. Acesso em: abr.2017.