UMA PROPOSTA DE ECOSSISTEMA DE *BIG DATA* PARA A ANÁLISE DE DADOS ABERTOS GOVERNAMENTAIS CONECTADOS

Marcio de Carvalho Victorino[*]
Marcelo Shiessl[**]
Edgard Costa Oliveira[***]
Edson Ishikawa[****]
Maristela Terto de Holanda[*****]
Marçal de Lima Hokama[******]

RESUMO O presente estudo trata da apresentação de uma proposta de Ecossistema de Big Data para dar suporte à análise de dados abertos governamentais conectados. O ambiente de Big Data caracteriza-se por um conjunto de dados de grande volume, grande variedade de formatos e com a necessidade de serem processados a uma velocidade adequada. No referido Ecossistema, o processamento de dados massivos se dá por meio do uso de novas abordagens das áreas de Ciência da Informação e Ciência da Computação, que envolvem tecnologias e processos para a coleta, representação, armazenamento e disseminação da informação. Utiliza-se um modelo de Arquitetura da Informação composto por princípios de usabilidade, metadados, tesauros, taxonomias e ontologias para organizar e representar esse enorme volume de dados e a respectiva semântica. Com a implantação do Ecossistema, pretende-se proporcionar ao usuário final consultar um grande volume de dados públicos das mais diversas áreas do governo; ao profissional da informação, identificar fontes de dados relevantes, a fim de preparar um ambiente apropriado à tomada de decisão, com base na análise e mineração de dados; e, ao gestor público, realizar as análises em busca de *insights* que possam ajudar no estabelecimento e monitoramento de políticas públicas eficazes.

Palavras-chave: Big Data. Ecossistema. Dados abertos. Dados conectados. Arquitetura da informação.

* Doutor em Ciência da Informação pela Universidade de Brasília, Brasil. Professor Substituto do Departamento de Ciência da Computação da Universidade de Brasília, Brasil.
E-mail: mcvictorino@uol.com.br.

** Doutor em Ciência da Informação pela Universidade de Brasília, Brasil. Analista de Risco Financeiro e Corporativo da Caixa Econômica Federal, Brasil. Membro do Grupo de Pesquisa EROIC.
E-mail: marcelo.schiessl@gmail.com.

*** Doutor em Ciência da Informação pela Universidade de Brasília, Brasil. Professor Adjunto da Universidade de Brasília, Brasil.
E-mail: ecosta@unb.br.

**** Doutor em Engenharia de Sistemas e Computação pela Universidade Federal do Rio de Janeiro, Brasil. Professor Adjunto do Departamento de Ciência da Computação da Universidade de Brasília, Brasil.
E-mail: ishikawa@cic.unb.br.

***** Doutora em Engenharia Elétrica pela Universidade Federal do Rio Grande do Norte, Brasil. Professora Adjunta do Departamento de Ciência da Computação da Universidade de Brasília, Brasil.
E-mail: mholanda@cic.unb.br.

****** Mestrando do Curso de Pós-Graduação em Computação Aplicada do Departamento de Ciência da Computação da Universidade de Brasília, Brasil. Administrador de Dados Corporativos do Exército Brasileiro, Brasil.
E-mail: lima@cds.eb.mil.br.

1 INTRODUÇÃO

Há alguns anos, o governo brasileiro vem demonstrando a intenção de tornar a sua administração o mais transparente possível por meio da publicação de informações de interesse da sociedade na *web*. Em 2006, a Controladoria-Geral da União – CGU – agora, Ministério da Transparência, Fiscalização e Controle – órgão responsável pelo controle interno do Governo Federal, em conjunto com o Ministério do Planejamento, Orçamento e Gestão – MPOG, estabeleceram a Portaria Interministerial CGU/MPOG n. 140, de 16 de março de 2006, que determina que os órgãos e entidades da Administração Pública Federal são responsáveis por manter nos seus respectivos sítios eletrônicos as informações detalhadas sobre

determinados aspectos, como, por exemplo, execução orçamentária, licitações, contratações, entre outros. Estas devem ser mantidas em páginas específicas, denominadas Páginas de Transparência Pública (CGU, 2006).

Segundo a CGU, as Páginas de Transparência têm como missão promover a visibilidade dos gastos públicos e incentivar o controle social para que as práticas da Administração Pública sejam pautadas pela legalidade e ética (CGU, 2006).

Em 2011, o Brasil passou a integrar a *Open Government Partnership* (OGP), uma instituição com o objetivo de fornecer uma plataforma internacional para tornar os governos mais abertos. A OGP (2011) destaca diversos benefícios das iniciativas de abertura de dados, tais como: a melhoria dos serviços públicos e mais compreensão das atividades governamentais; a gestão mais efetiva dos recursos públicos; o aumento da responsabilização e da prestação de contas; o aumento da integridade pública; a criação de comunidades mais seguras; e, uma maior participação do cidadão na gestão pública.

Neste sentido, além das Páginas de Transparência Pública que apresentam dados referentes às despesas realizadas por cada órgão e entidade da Administração Pública Federal, o Governo Federal também disponibiliza informações sobre a aplicação de recursos públicos, a partir da consolidação de milhões de dados oriundos de diversos órgãos federais relativos a Programas e Ações de Governo em um único sítio denominando Portal da Transparência (CGU, 2012). Esses ambientes disponibilizam um massivo volume de dados públicos estruturados, semiestruturados e não estruturados de interesse coletivo ou geral. Assim, torna-se um grande desafio a criação de aplicações capazes de gerar *insights* em uma velocidade apropriada a partir do enorme volume de dados nos mais variados formatos. Para lidar com a questão da velocidade, do volume e da variedade, tem-se um novo conceito de ambiente de tratamento da informação: o *Big Data*.

O processamento desse recurso informacional tem demandado o estudo de novas abordagens nas áreas de Ciência da Informação e Ciência da Computação, que envolvem tecnologias e processos para de coleta, representação, armazenamento e disseminação da informação. Ciente da complexidade deste problema, um grupo formado por pesquisadores das áreas de Ciência da Informação e Ciência da Computação vislumbrou a possibilidade de iniciar um projeto de pesquisa que pudesse estruturar um Ecossistema de *Big Data* para dar suporte à análise de dados abertos governamentais conectados.

Diante do exposto, o presente estudo tem por objetivo a apresentação da especificação, em um alto nível de abstração, de um Ecossistema de *Big Data* para dar suporte à produção e consumo de dados abertos governamentais de qualidade no âmbito do governo brasileiro. Este deverá possibilitar o armazenamento dos dados oriundos de diversas origens para serem tratados e, posteriormente, servirem de subsídio para a avaliação e o monitoramento de programas sociais, com o objetivo de apoiar o desenho e a gestão de políticas públicas.

2 DADOS ABERTOS GOVERNAMENTAIS CONECTADOS

Segundo a definição da *Open Knowledge International*, antes conhecida como *Open Knowledge Foundation*, os "dados são abertos quando qualquer pessoa pode livremente usá-los, reutilizá-los e redistribuí-los, estando sujeito a, no máximo, a exigência de creditar a sua autoria e compartilhar pela mesma licença" (TCU, 2015, p. 5).

Sobre tal questão, a *World Wide Web Consortium* (W3C), um consórcio internacional com a missão de conduzir a *web* ao seu potencial máximo por meio da criação de padrões e diretrizes que garantam sua evolução permanente – endossa a definição proposta por Eaves (2009, p. 1):

> Dados Abertos Governamentais são a publicação e a disseminação das informações do setor público na *web*, compartilhadas em formato bruto aberto, compreensíveis logicamente, de modo a permitir sua reutilização em aplicações digitais desenvolvidas pela sociedade.

Eaves (2009), especialista em políticas públicas e ativista dos dados abertos, propôs três leis que foram adotadas pelo W3C, quais sejam:

- Se o dado não pode ser encontrado e indexado na *web*, ele não existe;
- Se não estiver aberto e em formato compreensível por máquina, ele não pode ser reaproveitado; e
- Se algum dispositivo legal não permitir sua reaplicação, ele não é útil.

Em 08 de dezembro de 2007, 30 americanos defensores de dados abertos, representados por pesquisadores de organizações da sociedade civil e ativistas, se reuniram para desenvolver um conjunto de princípios para os dados abertos governamentais. O encontro, realizado em Sebastopol, Califórnia, Estados Unidos da América – EUA, foi projetado para desenvolver um entendimento mais robusto a respeito do porquê dos dados abertos governamentais serem essenciais para a democracia. Aquele grupo propôs um conjunto de oito princípios fundamentais para os dados abertos governamentais, a saber (TCU, 2015):

- Completos: todos os dados públicos estão disponíveis. Entende-se por dado público o dado que não está sujeito a limitações válidas de privacidade, segurança ou controle de acesso.
- Primários: os dados são apresentados tais como os coletados na fonte, com o maior nível de granularidade e sem agregação ou modificação.
- Atuais: os dados são disponibilizados tão rapidamente quanto necessária à preservação do seu valor.
- Acessíveis: os dados são disponibilizados para o maior alcance possível de usuários e para o maior conjunto possível de finalidades.
- Compreensíveis por máquinas: os dados são razoavelmente estruturados de modo a possibilitar processamento automatizado.
- Não discriminatórios: os dados são disponíveis para todos, sem exigência de requerimento ou cadastro.
- Não proprietários: os dados são disponíveis em formato sobre o qual nenhuma entidade detenha controle exclusivo; e
- Livres de licenças: os dados não estão sujeitos a nenhuma restrição de direito autoral, patente, propriedade intelectual ou segredo industrial. As restrições sensatas relacionadas à privacidade, segurança e privilégios de acesso devem ser permitidas.

Em 2015, o Tribunal de Contas da União – TCU – órgão responsável pelo controle externo do Governo Federal – publicou um documento no qual elenca cinco motivos para a abertura de dados na Administração Pública brasileira (TCU, 2015), quais sejam:

- Porque a sociedade exige mais transparência na gestão pública;
- Porque a própria sociedade pode contribuir com serviços inovadores ao cidadão;
- Porque ajuda a aprimorar a qualidade dos dados governamentais;
- Para viabilizar novos negócios; e
- Porque é obrigatório por Lei (Lei n. 12.527/2011).

Em 2006, Tim Berners-Lee publicou o documento "*Design Issues"* com uma subseção de *web* semântica exclusiva para a interoperabilidade entre dados. Aquele autor ressalta a importância da integração semântica desses dados, dando origem à área de dados conectados. O termo "dados conectados" (*linked data*) se refere a um conjunto de boas práticas para a publicação e conexão de dados estruturados na *web*, fazendo uso de padrões internacionais recomendados pelo W3C (ISOTANI; BITTENCOURT, 2015).

Com base na relevância da interoperabilidade dos dados abertos governamentais e privados, Tim Berners-Lee (2006) propôs o "Sistema de 5 Estrelas" – um sistema que classifica, por meio de estrelas, o grau de abertura dos dados; ou seja, quanto mais aberto, maior o número de estrelas para os dados e mais facilidade para ser enriquecido.

As cinco estrelas para os dados abertos são:

- "1 Estrela": disponível na *internet* (em qualquer formato; por exemplo, .PDF), desde que com licença aberta, para que seja considerado dado aberto;
- "2 Estrelas": disponível na *internet* de modo estruturado (por exemplo, em uma planilha MS-Excel);
- "3 Estrelas": disponível na *internet* de modo estruturado e em formato não proprietário (em uma planilha OpenOffice.

org ou *Comma Separated Values* – CSV em vez de MS-Excel);

- "4 Estrelas": seguindo todas as regras anteriores, mas dentro dos padrões estabelecidos pelo W3C (*Resource Description Framework* - RDF e *SPARQL Protocol and RDF Query Language* - SPARQL): uso de *Uniform Resource Locator* – URL para a identificação de coisas e propriedades, de modo que todos possam direcionar para suas publicações; e
- "5 Estrelas": todas as regras anteriores e mais a conexão de seus dados a outros dados, fornecendo um contexto.

Segundo Isotani e Bittencourt (2015), é aconselhável que os dados sejam abertos considerando no mínimo três estrelas. Porém, o ecossistema aqui proposto tem por objetivo atingir as cinco estrelas para os dados abertos governamentais conectados por meio de uma Arquitetura da Informação – AI, a fim de organizar e representar esse volume de dados massivo e a respectiva semântica.

3 ARQUITETURA DA INFORMAÇÃO

O termo "Arquitetura da Informação" – AI foi utilizado pela primeira vez em 1975, pelo arquiteto Wurman (2005), com base na importância da organização da informação para a sua compreensão tanto para os produtores quanto para os consumidores. O referido profissional afirma que os verdadeiros arquitetos da informação dão clareza ao que é complexo, tornando a informação compreensível a outros seres humanos.

De fato, não se tem uma definição precisa sobre o que é ou o que constitui uma AI, e entre os vários pesquisadores que escrevem sobre o assunto, é possível observar uma grande quantidade e diversidade de definições.

Para Davenport (1998), a AI é um conceito confuso, que pode abranger muitos significados alternativos. No entanto, na perspectiva ecológica, significa um guia para estruturar e localizar a informação dentro de uma organização.

Brancheau e Wetherbe (1986) afirmam que a AI consiste em um plano para modelagem dos requisitos informacionais de uma organização, que provê um modo de mapear as informações necessárias à própria organização, que se referem aos processos do negócio e à documentação de seus inter-relacionamentos.

Macedo (2005) afirma que a AI é uma metodologia de 'desenho' que se aplica a qualquer 'ambiente informacional', sendo este compreendido como um espaço localizado em um 'contexto', constituído por conteúdos em fluxo, que serve a uma comunidade de 'usuários'. Sua finalidade é, portanto, viabilizar o fluxo efetivo de informações por meio do desenho de 'ambientes informacionais'.

Na bibliografia atual é possível encontrar várias propostas de AI, dentre elas, pode-se citar Rosenfeld e Morville (2002); Morrogh(2002); Batley (2007); e Wodtke e Govella (2011). A proposta de Rosenfeld e Morville (2002) tornou-se um dos marcos mais importantes para a área de AI. Os autores propõem um modelo no qual a AI é representada como a interseção de contexto, conteúdo e usuários. No espaço informacional de uma organização é necessário conhecer os objetivos do negócio da organização (contexto), estar consciente da natureza e do volume de informações existentes e de sua taxa de crescimento (conteúdo), bem como, entender as necessidades e os processos de busca do público-alvo (usuários).

Rosenfeld e Morville (2002) apresentam uma visão direcionada quase que exclusivamente para o desenvolvimento se sites, no entanto, os recursos de AI utilizados se aplicam a quaisquer coleções de informações, dentre eles, esquemas de organização, rotulação e navegação de um sistema de informação.

Para esse trabalho é utilizada uma adaptação da AI proposta por Victorino (2011). Essa proposta é embasada nos mesmos princípios de Rosenfeld e Morville (2002), no entanto, apresenta uma extensão para ser utilizado em qualquer ambiente informacional. Em sua composição estão presentes recursos de usabilidade, metadados, tesauros, taxonomias e ontologias.

a) *Usabilidade*

De acordo com Bohmerwald (2005), os critérios de usabilidade fornecem parâmetros para medir a eficiência da *interface* e revela como se dá a interação usuário-sistema. Segundo Bevan

(1995), por "usabilidade" se entende a qualidade da interação dos usuários com uma determinada *interface* – qualidade associada aos seguintes princípios: facilidade de aprendizado; facilidade de lembrar como realizar uma tarefa após algum tempo; rapidez no desenvolvimento de tarefas; baixa taxa de erros; e, satisfação subjetiva do usuário.

b) Metadados

O metadado pode ser definido como o dado ou a informação sobre o dado. Normalmente, é utilizado para armazenar informações úteis à recuperação ou acesso à informação, devendo ser capaz de descrever ou servir de sumário para o conteúdo de determinada informação. O termo surgiu em 1995, por ocasião de um simpósio realizado em Dublin, Ohio, EUA, que deu origem à *Dublin Core Metadata Iniciative* – DCMI (2012).

c) Tesauro

O termo "tesauro" tem origem no dicionário analógico de Peter Mark Roger, intitulado *Thesaurus of English Words and Phrases*, publicado, pela primeira vez em Londres, em 1852 (GOMES, 1990). Sobre a questão, Gomes (1990, p. 16) aponta que o tesauro nada mais é do que "uma linguagem documentária dinâmica que contém termos relacionados semântica e logicamente, cobrindo de modo compreensivo um domínio do conhecimento".

Em suma, o tesauro é uma lista estruturada de termos, associada e empregada por analistas de informação e indexadores para descrever um documento com a desejada especificidade, em nível de entrada, além de permitir aos pesquisadores a recuperação da informação que procuram (CAVALCANTI, 1978).

d) Taxonomia

Carl Linnaeus é conhecido como o "Pai da Taxonomia", pois seu sistema para nomeação, ordenação e classificação de organismos é até hoje de grande valia. Sobre a questão, Campos e Gomes (2008, p. 1) afirmam: "Taxonomia é, por definição, classificação sistemática e está sendo conceituada no âmbito da Ciência da informação como ferramenta de organização intelectual".

Segundo Conway e Sligar (2002), não há uma definição consensual para o termo "taxonomia". Neste sentido, aquelas autoras distinguem três tipos de taxonomia, a saber: descritiva, navegacional e de vocabulário de gerenciamento de dados, e propõem o uso do termo para referenciar qualquer coleção classificada de elementos.

e) Ontologia

Historicamente, o termo "ontologia" tem origem no grego *ontos* (ser) e *logos* (tratado). O termo original é a palavra aristotélica "categoria" – utilizada para classificar alguma coisa. Aristóteles apresenta categorias que servem de base para classificar qualquer entidade, e introduz ainda o termo "*differentia*" para propriedades que distinguem diferentes espécies do mesmo gênero. A conhecida técnica de herança é o processo de mesclar *differentias*, definindo categorias por gênero (BAX; ALMEIDA, 2003).

Guarino (1998) ressalta o uso predominante de ontologias na Inteligência Artificial – IA, definindo-as como um artefato de engenharia constituído de um vocabulário específico, utilizado para descrever uma determinada realidade e um conjunto de suposições explícitas, relacionadas ao significado intencional das palavras do vocabulário.

Muitas definições foram apresentadas nas últimas décadas, mas, a mais citada, no contexto das áreas de Ciência da Informação e Ciência da Computação, tem por base a proposta de Gruber (1993) que se desdobra nas seguintes definições:

- Definição 1 — Gruber (1993) propôs que a ontologia é uma especificação de uma conceituação;
- Definição 2 — Borst (1997) complementou afirmando que a ontologia é uma especificação de uma conceituação compartilhada; e
- Definição 3 — Studer, Benjamins e Fensel (1998) combinaram as definições supramencionadas ao estabelecer que a ontologia é uma especificação explícita e formal de uma conceituação compartilhada.

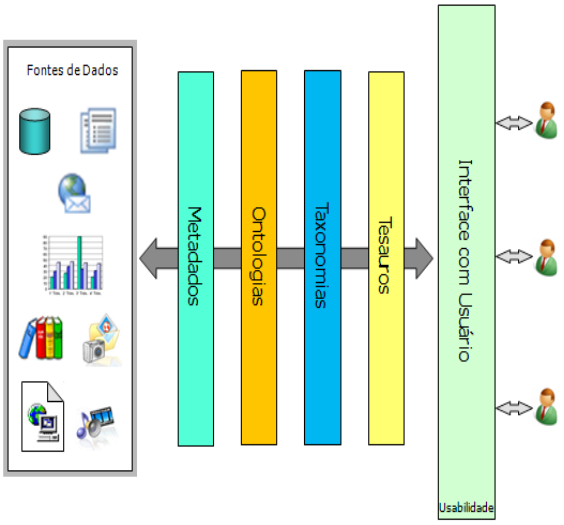
Uschold e Grüninger (1996) afirmam que a ontologia é o termo utilizado para se referir ao

entendimento compartilhado de um determinado domínio de interesse, que pode ser utilizado como uma estrutura unificada para solucionar vários tipos de problemas, entre os quais, aqueles relacionados ao compartilhamento do conhecimento e de interoperabilidade.

4 ORGANIZAÇÃO DA INFORMAÇÃO

No presente estudo, tem-se que o objetivo de uma AI é a organização e armazenagem dos dados estruturados, semiestruturados e não estruturados em repositórios informacionais (bancos de dados, sistemas de arquivos etc.) providos de consistência, compartilhamento, documentação, privacidade e recuperação eficaz de seus conteúdos. Neste sentido, a Figura 1, a seguir, apresenta a AI proposta por Victorino (2011), devidamente adaptada para a pesquisa em questão.

Figura 1 – Modelo conceitual de Arquitetura da Informação.



Fonte: Adaptado de Victorino (2011).

De acordo com a AI apresentada na Figura 1, para o acesso a uma determinada fonte de dados, o usuário interage com *interfaces* implementadas conforme os critérios de usabilidade. Os tesauros são utilizados para permitir ao usuário encontrar o termo que represente um determinado significado para aquilo que procura. As taxonomias navegacionais são utilizadas para permitir que os usuários leigos naveguem pelo conteúdo do repositório, sendo criadas levando em conta o comportamento do usuário. As ontologias permitem o aprimoramento da indexação das fontes de dados, por meio da representação semântica, e das buscas realizadas pelos usuários por meio da delimitação do contexto. E, por fim, os metadados descrevem o suporte e o conteúdo, servindo de índices para a recuperação da informação.

Neste ambiente, outro desafio importante a ser encarado consiste em criar aplicações capazes de gerar *insights*, compreensão de situações ou problemas complexos e a percepção dos elementos que levam a sua resolução, em uma velocidade apropriada a partir de um enorme volume de dados, disponibilizados em uma grande variedade de formatos. Para lidar com tais desafios – velocidade, volume e variedade – tem-se novos conceitos de ambiente de tratamento da informação, a saber: o *Big Data* e o *Big Data Analytics*.

5 BIG DATA

Estima-se que do início da civilização até 2003, a humanidade tenha criado cinco *exabytes* (10 bytes elevados a 18ª potência) de dados. Atualmente cria-se esse mesmo volume de dados a cada dois dias (SCHMIDT, 2010). Um estudo da *International Data Corporation* – IDC (GANTZ; REINSEL, 2011) indica que de 2012 até 2020, o volume de dado armazenado na *internet* deverá dobrar a cada dois anos.

Algumas explicações para tal fenômeno são: a drástica redução de preços para o armazenamento das informações; a explosão de aplicações disponíveis na *internet* (*e-commerce*); a popularização de sensores conectados – *internet* das coisas, pesquisas científicas – ao projeto genoma; e, as redes sociais (Facebook, Twitter etc.).

Tal cenário demanda soluções efetivas em termos de custos e formas inovadoras de tratamento da informação para uma melhor percepção e tomada de decisão. Uma das propostas emergentes para lidar com esse ambiente complexo é o *Big Data*.

Há várias definições, entendimentos e discussões para o termo "*Big Data*", e uma das mais aceitas é a definição de '3Vs', apresentada por Laney (2001). O "*Big Data*" é caracterizado por um conjunto de dados de grande 'volume', adquiridos em alta 'velocidade' e com informações de alta 'variedade' de formatos. Davenport (2014) ressalta que outros 'Vs' também já foram acrescentados à definição inicial de *Big Data*, quais sejam: 'veracidade' e 'valor'.

Davenport (2014) afirma que o *Big Data* é inegavelmente grande. Por outro lado, possui uma designação um tanto quanto inapropriada, pois se trata de um termo genérico para dados que não cabem em repositórios habituais. Segundo aquele autor, o *Big Data* se refere a dados massivamente volumosos para caberem em simples servidores, extremamente desestruturados para se ajustarem a bancos de dados com base em linhas e colunas de tabelas relacionais, e continuamente fluídos para caberem em estruturas estáticas de armazenagem.

Várias pesquisas têm sido desenvolvidas com o objetivo de conceber novas tecnologias para armazenar e processar tais dados, para disponibilizá-los para consultas e análises de suporte à decisão. Entre as tecnologias mais promissoras, encontram-se o *middleware* Hadoop e os sistemas gerenciadores de banco de dados NoSQL.

O Hadoop destaca-se por ter implementado o conceito apresentado pela empresa Google, denominado MapReduce (DEAN; SANJAY, 2008). É uma abordagem que busca dividir os problemas complexos do *Big Data* em pequenas unidades de trabalho e processá-las em paralelo.

Em uma rede de computadores onde o dado encontra-se distribuído pelos nós de computadores que compõem essa rede, o MapReduce pode ser dividido em dois estágios, a saber (DEAN; SANJAY, 2008):

- Passo de mapeamento: o nó mestre divide os dados em vários subconjuntos menores; um nó trabalhador processa um subconjunto de dados menor sob o controle de um rastreador de trabalho e armazena o resultado no sistema de arquivos local, onde um redutor será capaz de acessá-lo.
- Passo de redução: analisa e reúne os dados de entrada a partir das etapas de mapeamento. Pode haver múltiplas tarefas de redução para paralelizar o processamento. Estas são executadas nos nós trabalhadores sob o controle do rastreador de trabalho.

Indrawan-Santiago (2012) apresenta a seguinte classificação dos gerenciadores de banco de dados NoSQL, a partir do modelo de dados utilizado:

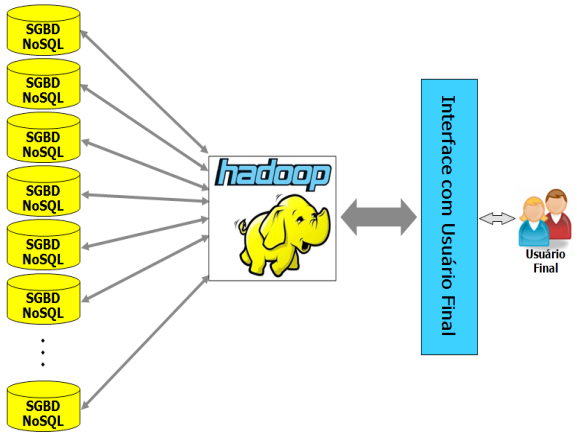
- chave-valor: os dados são armazenados sob a forma de pares chave/valor, de modo estruturado ou não estruturado. Cada uma das chaves é única, e os clientes atribuem ou solicitam os valores para cada chave.
- colunar: os bancos NoSQL colunares, ou orientados à coluna, conhecidos também como família de colunas, armazenam e processam os dados por coluna ao invés de linhas, como ocorre no banco de dados relacional.
- orientado a documentos: um banco NoSQL orientado a documentos faz uso do conceito de par chave/valor para o armazenamento de dados. Todavia, é imposta alguma estrutura em como o valor é armazenado. Em comparação com os bancos de dados chave-valor, isso provê mais informação sobre a estrutura, além de suportar estruturas mais complexas.
- orientado a grafos: nesta categoria de banco NoSQL, os dados são representados como uma rede de nós conectados por arestas, o que permite a determinação e qualificação da conectividade entre as entidades.

Indrawan-Santiago (2012) também realiza uma reflexão sobre os diversos tipos de banco de dados NoSQL apresentando as principais vantagens e desvantagens destes quando comparados aos relacionais. As principais vantagens apontadas por aquela autora são: a flexibilidade de suas estruturas, a alta escalabilidade horizontal, o suporte a dados não estruturados e o processamento distribuído. Tais características tornam os bancos de dados NoSQL excelentes dispositivos de persistência para o ambiente de *Big Data*.

O *framework* Hadoop e os bancos de dados NoSQL são recursos vitais em um ambiente

de *Big Data*. Neste sentido, a Figura 2, a seguir, apresenta a arquitetura convencional de *Big Data* de Davenport (2014), com algumas adaptações.

Figura 2 – Arquitetura convencional de um ambiente de *Big Data*.



Fonte: Adaptado de Davenport (2014).

Na arquitetura apresentada na Figura 2, os dados são persistidos em centenas ou milhares de bancos de dados NoSQL. O Hadoop age como um *middleware* que, após mapear os dados distribuídos e aplicar as transformações ou regras de negócio (MapReduce), consolida o resultado e apresenta a resposta ao usuário final. É preciso destacar que o Hadoop é um *framework* de computação distribuída implementado em Java, voltado para *clusters* e processamento de grandes massas de dados. Em verdade, trata-se de um conjunto de tecnologias.

6 BIG DATA ANALYTICS

O *Big Data Analytics* pode ser interpretado como procedimentos complexos que são executados em larga escala sobre grandes repositórios de dados, cujo objetivo é a extração de conhecimento útil mantido em tais repositórios (CUZZOCREA, 2013). Em outras palavras, é a aplicação de técnicas analíticas avançadas a grandes conjuntos de dados.

Davenport e Kim (2013) classificam a análise em ambientes de *Big Data* de acordo com seus métodos e processos. Aqueles autores propõem três tipos de análise, a saber: descritiva, preditiva e prescritiva, levando-se em conta os métodos utilizados. E ainda, tem-se outros dois tipos de análise – qualitativa e quantitativa – levando-se em conta os processos utilizados.

A análise descritiva envolve as ações de coleta, organização, tabulação e apresentação de dados para a exposição das características do que está sendo estudado, sendo denominada "elaboração de relatório" ou "resumo de dados". Consiste de um recurso que pode ser muito útil, mas não explica os resultados ou as ocorrências, nem indica o que pode acontecer no futuro.

Por outro lado, a análise preditiva vai além da mera descrição das características dos dados e das relações entre as variáveis, uma vez que faz uso dos dados do passado para prever o futuro. Em tal análise, primeiro são identificadas as associações entre as variáveis e, em seguida, faz-se a previsão da probabilidade da ocorrência de um fenômeno, levando-se em conta as relações identificadas.

Já a análise prescritiva, por meio da inclusão de métodos como projeto experimental e otimização, se estende ainda mais. Tal como a receita de um médico, a análise prescritiva sugere um curso de ação. Nesta, o projeto experimental tenta responder às perguntas sobre porque algo aconteceu, por meio de experimentos, e a otimização tenta descobrir o nível ideal de determinada variável em suas relações com outra.

Finalmente, a análise qualitativa tem por objetivo promover a compreensão das razões e motivações subjacentes a um fenômeno por meio da observação de um pequeno número de casos representativos, enquanto que a análise quantitativa almeja a investigação empírica sistemática de um fenômeno por meio da observação de um grande número de casos e posterior tratamento, fazendo uso de técnicas estatísticas, matemáticas ou computacionais.

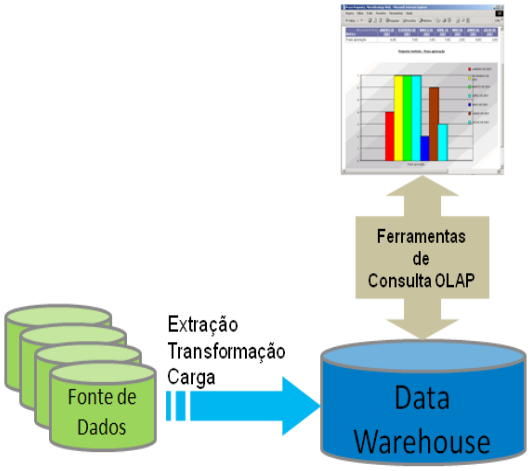
a) Arquitetura On Line Analytical Processing em Big Data Analytics

Os Sistemas de Apoio à Decisão – SAD, antes do advento do *Big Data*, normalmente eram organizados em uma arquitetura *On Line Analytical Processing* – OLAP, que possui um repositório de dados multidimensional,

denominado *Data Warehouse* – DW, capaz de armazenar os dados oriundos de diversas fontes. Para compor um DW, os dados, oriundos de sistemas de informação transacionais, passam por um processo de extração, transformação e carga (em inglês, *Extraction*, *Transformation*, *Load* – ETL), para que possam ser analisados de modo integrado.

Diante do exposto, a Figura 3, a seguir, apresenta os componentes de uma arquitetura OLAP sugeridos por Kimball e Ross (2013) para um ambiente de apoio à decisão: fontes de dados transacionais; camada de ETL; DW; e, servidor de relatórios analíticos que dão suporte completo às consultas *ad hoc*, onde o decisor pode navegar pelos dados organizados dimensionalmente em um DW, a fim de gerar planilhas ou gráficos, sem que seja necessária a criação de uma linha de código.

Figura 3 – Arquitetura *On Line Analytical Processing* convencional.



Fonte: Adaptado de Kimball e Ross (2013).

Na arquitetura OLAP apresentada na Figura 3, o DW é, portanto, um repositório de dados corporativos, onde os dados obtidos de sistemas-fonte são devidamente tratados e, posteriormente, depositados em bancos de dados informacionais, que oferecem um enfoque histórico para permitir um suporte efetivo à decisão.

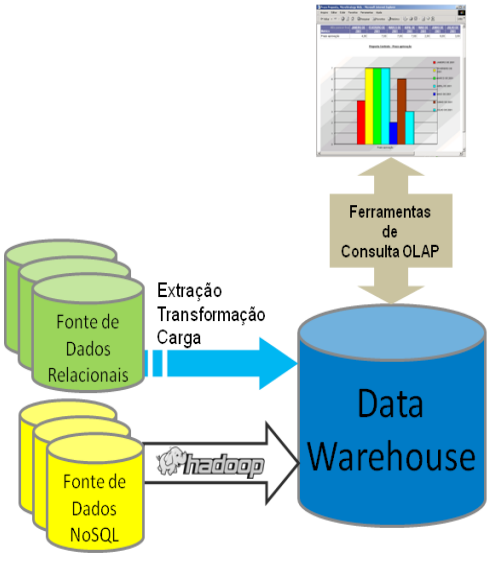
No entanto, Kimball e Ross (2013) afirmam que bancos de dados relacionais e a linguagem de consulta SQL, muito utilizados em DWs convencionais, não são capazes de analisar a grande quantidade de dados de um *Big Data*.

Por outro lado, Indrawan-Santiago (2012) ressalta que os bancos de dados NoSQL, muito utilizados em ambientes de *Big Data*, foram concebidos para dar suporte a processos operativos que manipulam um grande volume de dados em diversos formatos em um tempo aceitável, e não para fazer análises mais elaboradas, a fim de proporcionar suporte à decisão. Outro aspecto importante que deve ser observado é que existem pouquíssimas ferramentas de consultas analíticas a repositórios de dados disponíveis para o Hadoop – principal *framework* utilizado em ambiente de *Big Data*.

O grande desafio consiste em disponibilizar uma arquitetura para que os dados estruturados e não estruturados de um ambiente de *Big Data* possam ser utilizados em conjunto, a fim de dar suporte aos processos de tomada de decisão. Neste sentido, Davenport (2014) apresenta uma proposta para fundir as duas tecnologias – OLAP e *Big Data* – em um ambiente integrado.

A Figura 4, a seguir, apresenta uma extensão da arquitetura OLAP convencional, segundo Davenport (2014), a fim de proporcionar o seu uso em ambiente de *Big Data*.

Figura 4 – Arquitetura *On Line Analytical Processing* estendida



Fonte: Adaptado de Davenport (2014).

Na arquitetura apresentada na Figura 4, as fontes de dados operativas encontram-se nos mais variados formatos, representadas por bancos de dados relacionais e NoSQL. Os dados armazenados em bancos de dados relacionais passam por um processo ETL e são armazenados no DW, tal como em uma arquitetura OLAP convencional. Já os dados armazenados em bancos de dados NoSQL passam por um processo ETL executado pelo *framework* Hadoop e são armazenados no mesmo DW. O Hadoop também pode funcionar como ferramenta de apresentação de dados para as consultas operativas.

Nesta arquitetura, todos os dados analíticos são armazenados em um DW, proporcionando uma análise integrada a partir do cruzamento de informações de diversas áreas, gerando o resultado por meio de relatórios OLAP. Vale ressaltar que tal arquitetura é limitada, pois o ambiente OLAP atual não é capaz de suportar o volume e variedade de dados de um ambiente de *Big Data* (KIMBALL, ROSS; 2013).

Uma solução mais completa seria a geração de relatórios analíticos diretamente a partir dos repositórios de dados que compõem o *Big Data*, por meio de ferramentas de consultas que fazem parte do *framework* Hadoop, tendo em vista que este último não possui limitações quanto a acesso a fontes de dados heterogêneas e distribuídas. Entretanto, as ferramentas de análise disponível para o Hadoop não possuem a maturidade das ferramentas OLAP.

Diante do exposto, após o estudo minucioso das várias opções arquiteturais de *Big Data Analytics*, chegou-se à conclusão que, inicialmente, o ecossistema aqui descrito é passível para a adoção da arquitetura apresentada na Figura 4. No entanto, devido ao fato desta ser limitada, pretende-se, em um segundo momento, migrar para uma arquitetura mais abrangente com a eliminação do repositório do DW e a geração de relatórios analíticos diretamente pelo Hadoop. Para tanto, faz-se importante que as ferramentas de consultas analíticas do Hadoop estejam em condições de substituir as ferramentas de relatório OLAP convencionais.

b) *Mineração de Dados em Big Data Analytics*

A mineração de dados tem por objetivo o emprego de técnicas de aprendizado computacional, a fim de analisar e extrair automaticamente o conhecimento de grandes volumes de dados. No caso do *Big Data Analytics*, o processo de mineração de dados pode ser realizado de acordo com o modelo de referência *Cross Industry Standard Process for Data Mining* – CRISP-MD (SHEARER, 2000), organizado nas fases que se seguem:

- O Entendimento do Negócio (*Business Understanding*): foca o entendimento dos objetivos e requisitos do projeto, da perspectiva do domínio, a relevância do conhecimento prévio e os objetivos do usuário final;
- O Entendimento dos Dados (*Data Understanding*): realiza a coleta inicial de dados, descreve e explora os dados e verifica a qualidade dos dados;
- O Pré-Processamento dos Dados (*Data Preparation*): consiste na seleção de atributos, limpeza, construção, integração e formatação dos dados de entrada;
- A Modelagem (*Modeling*): seleciona modelos e parâmetros, com o uso direcionado para a obtenção de *insights*.
- A Avaliação (*Evaluation*): avalia, do ponto de vista de análise dos dados, a qualidade dos modelos obtidos, além de verificar se os objetivos do negócio foram atingidos conforme os critérios de sucesso adotados.
- A Implantação (*Deployment*): incorpora o modelo selecionado ao processo de tomada de decisão da organização.

7 ECOSSISTEMA DE *BIG DATA*

Existem diversas definições para o Ecossistema de *Big Data*. Shin e Choi (2015), por exemplo, apresentam uma definição em um contexto social mais amplo, como um ecossistema ecológico, que compreende as relações que envolvem os seguintes aspectos: tecnologia, governo, indústria, mercados, usuários e sociedade. No referido ecossistema, são examinados os efeitos do *Big Data* em todos os setores envolvidos. Mantendo o sentido biológico da interação entre os diversos componentes do cenário, mas em uma escala menor, Demchenko, Laat e Membrey (2014) definem um Ecossistema de *Big Data* como um complexo de facilidades técnicas e componentes

construídos em volta de uma origem de dados específica e sua aplicação – o complexo de componentes inter-relacionados é voltado para o armazenamento, processamento, visualização e entrega dos resultados a partir do *Big Data*. O ecossistema em questão ainda compreende, além do próprio *Big Data*, as seguintes categorias de componentes arquiteturais:

- Modelos e estruturas de dados: conforme Demchenko, Laat e Membrey (2014), os diversos estágios da transformação do *Big Data* requerem diferentes estruturas de dados, modelos e formatos, incluindo a possibilidade de processar tanto dados estruturados como desestruturados. É possível que as estruturas de dados e modelos correspondentes sofram modificações durante os diferentes estágios de processamento de dados. Todavia, é importante manter a ligação entre essas estruturas.
- Arquitetura de *Big Data*: é constituída pelo conjunto de tecnologias e componentes para o processamento e a análise do *Big Data*. Aqui Demchenko, Laat e Membrey (2014) ressaltam dois grupos de tecnologias principais, que denominam de *Big Data Analytics Infrastructure* – BDAI, quais sejam: a arquitetura geral que compreende as tecnologias e os componentes para o armazenamento, a computação, a rede, os dispositivos e o suporte operacional ao *Big Data*; e, a arquitetura de análise e processamento, que compreende as ferramentas de análise e processamento de dados, além da apresentação e visualização.
- Gerenciamento do ciclo de vida do *Big Data* (ou fluxo de transformação dos dados): Demchenko, Gruengard e Klous (2014) ressaltam a necessidade da utilização de métodos científicos para a obtenção dos benefícios das novas oportunidades de coleta e mineração de dados, a fim de lograr a informação desejada. O ciclo de vida requer o armazenamento e a preservação de dados em todos os estágios, com o intuito de possibilitar o reuso/redirecionamento e pesquisa/*analytics* nos dados processados e resultados publicados. Todavia, tal fato somente é possível se estiverem implementadas a identificação completa, a referência cruzada e a ligação dos dados. A integridade dos dados, do controle de acesso e da auditoria deve ser suportada durante todo o ciclo de vida dos dados.
- Infraestrutura de segurança do *Big Data*: compreende o conjunto necessário de componentes e políticas para prover controle no acesso aos dados e um ambiente de processamento seguro.

8 METODOLOGIA

O presente trabalho descreve o resultado da primeira etapa de uma pesquisa em andamento. As etapas consideradas para a consecução desta pesquisa completa foram agrupadas em duas partes distintas: a primeira parte, descrita neste trabalho, consistiu em conceber um modelo em alto nível de abstração de um Ecossistema de Big Data; e a segunda parte, em fase de desenvolvimento, foca a materialização desse modelo e a disponibilização de toda a infraestrutura necessária. As metodologias utilizadas para cada etapa são levantamento bibliográfico e sistemas flexíveis, respectivamente.

a) Levantamento Bibliográfico

O levantamento bibliográfico abrangeu textos que abordam as temáticas de Big Data, ecossistema, dados abertos, dados conectados, arquitetura da informação, arquitetura OLAP, mineração de dados e indexação semântica, entre outros, nas áreas da Ciência da Informação e Ciência da Computação.

Adotou-se o portal de periódicos disponibilizado pela Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Capes (CAPES, 2000) como principal fonte de informação sobre o tema, pois concentra uma grande quantidade de outros portais desde o ano de 2000. Além disso, outras buscas foram feitas diretamente, via navegador, em sites de instituições de ensino, repositórios públicos — como CiteSeerX (http://citeseerx.ist.psu.edu/index) —, Google Acadêmico (https://scholar.google.com.br/) e outros recursos disponíveis na *web* como complemento às bases mencionadas.

A partir deste levantamento, foi possível descrever a especificação, em um alto nível de abstração, de um Ecossistema de Big Data para dar suporte à produção e consumo de dados abertos governamentais de qualidade no âmbito do governo brasileiro.

b) Metodologia de Sistemas Flexíveis

A Metodologia de Sistemas Flexíveis (SSM – *Soft System Methodology*), desenvolvida na década de 60 pela equipe de Peter Checkland, é baseada no pensamento sistêmico. Ela enxerga o domínio do problema de forma holística, ao invés de enxergar de maneira reducionista, reconhecendo que as partes do sistema estão interconectadas, o que faz com que uma mudança em uma parte do sistema afete outras partes. Não obstante, o pensamento sistêmico reconhece que um problema em um domínio é apenas um subsistema de outros sistemas maiores. Dessa forma as mudanças podem afetar outros sistemas também (CHECKLAND, 1981).

A Metodologia de Sistemas Flexíveis possui sete etapas distintas:

- Estágio 1: situação-problema não estruturada;
- Estágio 2: situação-problema estruturada;
- Estágio 3: definições fundamentais dos sistemas relevantes;
- Estágio 4: construção de modelos conceituais;
- Estágio 5: comparação dos modelos conceituais (4) com a realidade (2);
- Estágio 6: identificação das mudanças desejáveis e possíveis;
- Estágio 7: ações para melhorar a situação-problema.

A SSM pode ser aplicada em problemas não-estruturados, na definição problemática de objetivos, em sistemas sociais, bem como nas disciplinas de Biologia, Ecologia, Economia, Demografia, Gestão, Engenharia, dentre outras (MAUAD *et al*., 2003). A metodologia utiliza uma abordagem holística para resolver problemas, os quais não podem ser resolvidos pela abordagem tradicional reducionista, com o fluxo da lógica baseada em indagações.

Costa (2003) constatou que o uso da SSM como Metodologia de pesquisa tem sido, via de regra, em pesquisa aplicada. A autora afirma que a utilização da SSM em projetos de pesquisa na Ciência da Informação pode contribuir para a discussão de questões típicas da área.

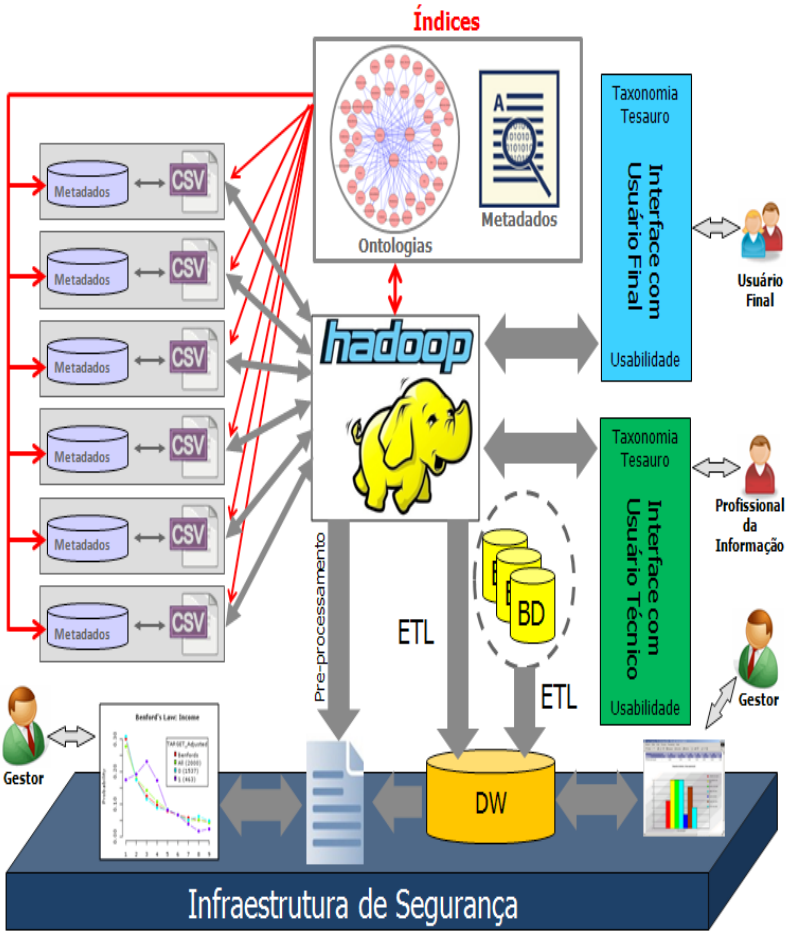
9 PROPOSTA DE ECOSSISTEMA DE *BIG DATA* PARA A ANÁLISE DE DADOS ABERTOS GOVERNAMENTAIS CONECTADOS

O primeiro passo para estruturar a proposta de um Ecossistema de *Big Data* para dar suporte à análise de dados abertos governamentais conectados consistiu da abstração de uma arquitetura de *Big Data* capaz de oferecer as funcionalidades requeridas. Este teve início pela análise da arquitetura convencional de *Big Data* apresentada por Davenport (2014). O objetivo era estendê-la, de modo a contemplar os elementos de um Ecossistema de *Big Data* elencados por Demchenko, Laat e Membrey (2014), bem como as regras para dados abertos conectados com grau de abertura "5 Estrelas" elencadas por Isotani e Bittencourt (2015). A Figura 2 apresenta essa arquitetura convencional.

Os principais aspectos observados para a extensão da arquitetura apresentada na Figura 2 foram a indexação e descrição semântica das fontes de informação por meio da integração da AI proposta por Victorino (2011), além da estratégia de indexação semântica da informação proposta por Schiessl (2015). A utilização de tais recursos tem por objetivo lidar com dados massivos distribuídos e implementar os requisitos obrigatórios para que os dados disponibilizados no ecossistema sejam considerados dados abertos conectados com grau de abertura "5 Estrelas". Neste sentido, a Figura 5, a seguir, apresenta a arquitetura estendida.

Na Figura 5 é possível observar que as fontes de dados são disponibilizadas no formato CSV, a fim de alcançar o grau de abertura "3 Estrelas". Para o alcance do grau de abertura "4 Estrelas", as fontes de dados serão endereçadas por meio de URLs, sendo escolhidos os padrões e as tecnologias estabelecidos ou reconhecidos pela W3C para a criação de metadados, tesauros, taxonomias e ontologias, com destaque para os que se seguem: XML (*eXtensible Markup Language*), OWL (*Web Ontology Language*), RDF e SPARQL. Finalmente, para alcançar o grau de abertura "5 Estrelas", os itens metadados, tesauros, taxonomias e ontologias auxiliarão na contextualizarão e indexação das informações do ecossistema.

Figura 5 – Arquitetura de um ambiente de *Big Data Analytics* com dados abertos conectados para o ecossistema proposto.



Fonte: Elaborado pelos autores

Fazendo uso, como referência, do Ecossistema de *Big Data* proposto por Demchenko, Laat e Membrey (2014), o ecossistema aqui proposto disponibilizará os seguintes componentes arquiteturais:

- Modelos e estruturas de dados: apesar dos diversos estágios da transformação do *Big Data* requererem diferentes estruturas de dados, modelos e formatos, incluindo a possibilidade de processar tanto dados estruturados como não estruturados, no presente caso, a informação será armazenada no ecossistema no formato CSV, devido às restrições das regras de abertura de dados. No entanto, é importante manter a ligação entre os arquivos CSV e as fontes de geração desses dados, que são os sistemas de informação dos órgãos da Administração Pública brasileira – ligação que será mantida por meio de metadados que poderão descrever, entre outros aspectos, qual órgão gerou os dados, a data de geração, o conteúdo e o suporte da fonte original dos dados.
- Arquitetura de *Big Data*: conforme apresentado na Figura 5, a arquitetura proposta prevê o uso de arquivos CSV fornecidos pelos órgãos públicos, arquivos XML ou bancos de dados XML para a armazenagem dos metadados, arquivos OWL para a armazenagem das ontologias, banco de dados relacionais modelados dimensionalmente para a armazenagem

dos dados do DW que serão utilizados para o apoio à decisão, e arquivos proprietários para o armazenamento de dados que serão minerados. O Hadoop será utilizado para fornecer processamento distribuído e serviços de *clusterização*. Para a análise *ad hoc* e geração de *dashboards* para o apoio à decisão, serão utilizadas, inicialmente, as ferramentas OLAP, como, por exemplo, as ferramentas componentes do *software* Pentaho. Para a mineração de dados, quando do emprego de técnicas de aprendizado computacional para analisar e extrair automaticamente o conhecimento dos dados, serão utilizadas ferramentas específicas (o Weka ou a ferramenta R, por exemplo).

- Gerenciamento do ciclo de vida do *Big Data*: é o componente mais complexo do ecossistema, haja vista requerer o armazenamento e a preservação de dados em todos os estágios, possibilitando o reuso/redirecionamento e a pesquisa/análise.

Conforme o exposto na Figura 5, apesar da publicação dos dados do ecossistema ser no formato padrão CSV, tem-se a necessidade da manutenção da referência às suas fontes originais. Assim, para que o ecossistema funcione a contento, é preciso o estabelecimento de políticas de publicação de dados nos órgãos da Administração Pública, prevendo formatos adequados e metadados mínimos. Após a validação da publicação dos dados abertos no ecossistema, respeitando os formatos e as descrições concernentes, as fontes e os metadados serão indexados por meio de palavras-chave e ontologias (indexação semântica), a fim de facilitar o processo de recuperação da informação.

Os dados poderão ser apresentados aos usuários finais após serem processados pelo Hadoop. Para facilitar a recuperação da informação, a camada de apresentação será composta por *interfaces* implementadas segundo os princípios de usabilidade. Esta também encapsulará tesauros e taxonomias navegacionais, a fim de enriquecerem as consultas à informação – consultas que farão uso das ontologias e metadados, que estão disponíveis como índices, proporcionando resultados mais próximos às necessidades do usuário final.

Por outro lado, o referido ambiente proporcionará aos profissionais da informação a possibilidade de construir aplicações mais sofisticadas que proporcionem análises complexas por parte dos usuários finais, como, por exemplo, uma análise OLAP por meio de consultas *ad hoc* ou *dashboards* e mineração de dados em busca de *insights*.

Para a análise OLAP, o primeiro passo consiste em fazer a ETL das fontes operativas para armazená-los em um DW. O profissional da informação poderá identificar tais fontes por meio de tesauros e taxonomias navegacionais encapsulados na camada de apresentação, além de ontologias e metadados, disponibilizados como índices e acessíveis por meio do Hadoop, para a descrição e contextualização das informações e estruturas. Após a identificação das fontes de interesse, o próprio Hadoop pode ser usado para a execução da ETL, transportando os dados dos arquivos CSV para o DW. No caso de fontes externas ao ambiente, é possível fazer uso de ferramentas convencionais de ETL para a carga do DW. Após finalizada a carga do DW, as ferramentas de análise podem ser utilizadas para a apresentação dos dados. Vale destacar que com a evolução das ferramentas de relatórios analíticos do *framework* Hadoop, o repositório do DW não se tornará mais necessário, uma vez que o Hadoop gerará diretamente os relatórios analíticos requeridos pelos usuários finais.

No que tange à mineração de dados, os primeiros passos são o entendimento do negócio e dos dados. O profissional da informação poderá executar essas tarefas por meio de tesauros e taxonomias navegacionais encapsulados na camada de apresentação, além de ontologias e metadados, disponibilizados como índices e acessíveis por meio do Hadoop, para a descrição e contextualização da área de negócio e respectivas fontes de dados. Após a identificação das fontes de interesse, o próprio Hadoop pode ser utilizado para a execução do pré-processamento dos dados, transportando os dados dos arquivos CSV para o formato proprietário da ferramenta de mineração a ser utilizada. No caso de dados já armazenados no DW, o pré-processamento

é mais simples, bastando transportá-los para o ambiente de mineração, pois, a princípio, as transformações dos dados necessárias já foram executadas anteriormente pela operação de ETL.

- Infraestrutura de segurança do *Big Data*: a segurança da informação prevista para o ecossistema tem como objetivo proteger os dados e seus respectivos refinamentos, para que o ecossistema atinja seus objetivos. Ela tem início com a proteção do dado na sua origem e permeia todo o ciclo da informação até o seu público alvo, devendo ser integrada ao processo corporativo da segurança da informação das organizações geradoras de dados envolvidas. Apesar do ecossistema tratar os dados abertos, tem-se a necessidade de se prover o controle do acesso aos dados e um ambiente de processamento seguro, haja vista que após a coleta, pretende-se agregar inteligência a este conjunto de dados. Assim, é possível a necessidade de restrição do acesso a tais *insights*. E também ocorre a preocupação de privacidade para garantir os direitos individuais constitucionais. Outro aspecto importante a ser analisado é a necessidade e viabilidade de cópias de segurança de parte dos dados, pois, mesmo sendo possível refazer o tratamento dos dados em uma situação de perda de informações, o tempo necessário empregado na referida atividade pode ser impeditivo. Portanto, tem-se aí as questões inerentes à confidencialidade, integridade, disponibilidade e autenticidade – características básicas da segurança da informação que devem ser preservadas.

10 CONSIDERAÇÕES FINAIS

Diante do exposto, não se pode duvidar que a divulgação de dados governamentais de forma aberta e conectada incrementa a transparência da administração pública e pode proporcionar inúmeros benefícios aos governos e cidadãos. No entanto, não é uma tarefa simples processar o volume massivo dos dados gerados pelo governo brasileiro, em uma enorme variedade de formatos a uma velocidade apropriada, a fim de gerar *insights* úteis a gestores públicos, cidadãos e organizações interessadas.

Neste ínterim, a presente pesquisa apresentou uma abordagem integrada para a criação de um Ecossistema de *Big Data* para a análise de dados abertos governamentais conectados, com base em princípios e técnicas das áreas de Ciência da Informação e Ciência da Computação, que envolvem tecnologias e processos de coleta, representação, armazenamento e disseminação da informação.

O ecossistema proposto tem por objetivo disponibilizar dados compatíveis com o nível "5 Estrelas" para os dados abertos governamentais conectados por meio de uma AI composta por princípios de usabilidade, metadados, tesauros, taxonomias e ontologias, a fim de organizar e representar o volume de dados massivo e a respectiva semântica.

Para dar continuidade a esta pesquisa, a equipe responsável, composta por pesquisadores da área de Ciência da Informação e Ciência da Computação, planeja conceber uma arquitetura de software capaz de dar suporte ao ecossistema apresentado a fim de proporcionar a sua materialização.

Pretende-se, com o referido ecossistema, proporcionar ao usuário final a consulta de um grande volume de dados públicos das mais diversas áreas do governo; ao profissional da informação, identificar fontes relevantes para preparar um ambiente apropriado à tomada de decisão, com base na análise e mineração de dados; ao gestor público, realizar análises em busca de *insights* que possam ajudar no estabelecimento de políticas públicas eficazes, proporcionando o emprego racional de recursos públicos para o desenvolvimento do País; além de incentivar a população em geral no acompanhamento das políticas públicas por meio do acesso irrestrito, no ambiente *web*, a dados abertos governamentais conectados consistentes.

Artigo recebido em 26/06/2016 e aceito para publicação em 23/01/2017

A PROPOSAL FOR BIG DATA ECOSYSTEM FOR THE GOVERNMENT LINKED OPEN DATA ANALYSIS

ABSTRACT *The present study proposes a Big Data Ecosystem to support the analysis of government linked open data. Big Data environments are characterized by large volume of data, in a wide variety of formats, which require appropriate velocity processing. In the Ecosystem proposed in this study, the processing of massive volumes of data is done using new approaches in Information Science and Computer Science, which involves technologies and processes for the collection, representation, storage and dissemination of information. An Information Architecture model, composed of usability principles, metadata, thesaurus, taxonomies and ontologies is used to organize and represent these enormous volumes of data and the respective semantics. With the implementation of the Ecosystem, we intend to provide the end user with the means to consult a large volume of public data from the most diverse areas of government. This aids information professionals in identifying sources of relevant data, to prepare an appropriate environment for making decisions, based on the analysis and data mining, and helps public managers carry out analyses in search of insights that can support them in establishing and monitoring public policies efficiently.*

Keywords: *Big Data. Ecosystem. Open data. Linked data. Information architecture.*

REFERÊNCIAS

BATLEY, S. **Information architecture for information professionals**. Elsevier, 2007.

BAX, M. P.; ALMEIDA, M. B. Uma visão geral sobre ontologias: pesquisa sobre definições, tipos, aplicações, métodos de avaliação e de construção. **Ciência da Informação**, Brasília, v. 32, n. 3, p. 7-20, set./dez. 2003. Disponível em: <http://www.scielo.br/pdf/ci/v32n3/19019.pdf>. Acesso em: 15 fev. 2016.

BERNERS-LEE, T. Linked data-design issues. **W3C**, 2006. Disponível em: < https://www.w3.org/DesignIssues/LinkedData.html>. Acesso em: 15 jan. 2016.

BEVAN, N. Usability is quality of use. In: 6TH INTERNATIONAL CONFERENCE ON HUMAN COMPUTER INTERACTION, Yokohama, v. 20, p. 349-354, jul. 1995. Disponível em: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.87.7123&rep=rep1&type=pdf> Acesso em: 15 out. 2015.

BOHMERWALD, P. Uma proposta metodológica para avaliação de bibliotecas digitais: usabilidade e comportamento da busca por informação na Biblioteca Digital da PUC/Minas. **Ciência da Informação**, v. 34, n. 1, p. 95-103, 2005. Disponível em: <http://www.bibliotecadigital.ufmg.br/dspace/bitstream/handle/1843/LHLS-69XPCF/mestrado___paula_bohmerwald.pdf?sequence=1>. Acesso em: 20 jan. 2016.

BORST, W. N. **Construction of engineering ontologies for knowledge sharing and reuse**. 1997. Tese (Doutorado) — Institute for Telematica and Information Technology, University of Twente, Enschede, The Netherlands. Disponível em: <http://doc.utwente.nl/17864/>. Acesso em: 10 maio 2016.

BRANCHEAU, J. C.; WETHERBE, J. C. Information Architectures: methods and practice. **Information Processing & Management**, v. 22, n. 6, p. 453-463, 1986.

CAMPOS, M. L. A; GOMES H. E. Taxonomia e classificação: a categorização como princípio. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 8, 2007, Salvador, Anais. Salvador: ANCIB, 2007. Disponível em: < http://www.enancib.ppgci.ufba.br/artigos/GT2--101.pdf>. Acesso em: 10 maio 2016.

FUNDAÇÃO COORDENAÇÃO DE APERFEIÇOAMENTO DE PESSOAL DE NÍVEL SUPERIOR - CAPES, **Portal de Periódicos**, 2000.

Disponível em:< http://www.periodicos.capes.gov.br/>. Acesso em: 25 abr. 2016.

CAVALCANTI, C. R. **Indexação e tesauro**: metodologia e técnica. Brasília: ABDF, 1978.

CHECKLAND, P. B. **Systems Thinkings, Systems Practice**. Chichester: Wiley, 1981.

CONTROLADORIA-GERAL DA UNIÃO – CGU. **Páginas da Transparência Pública**. Brasília, 2006. Disponível em: <http://www3.transparencia.gov.br/>. Acesso em: 28 jul. 2015.

_____. **Portal da Transparência**. Brasília, 2012. Disponível em: <http://www.portaldatransparencia.gov.br//>. Acesso em: 05 fev. 2017.

CONWAY, S.; SLIGAR, C. **Unlocking knowledge assets**. Washington: Microsoft Press, 2002.

COSTA, S. M. S. Metodologia de sistemas flexíveis aplicada a estudos em cência da informação: uma experiência pedagógica. **Transinformação**, Campinas, v. 15, n. 2, p. 259-271, maio/ago., 2003.

CUZZOCREA, A. Analytics over Big Data: exploring the convergence of DataWarehousing, OLAP and data-intensive cloud infrastructures. In: COMPUTER SOFTWARE AND APPLICATIONS CONFERENCE – COMPSAC, IEEE 37th Annual, 2013. p. 481-483.

DAVENPORT, T. H. **Ecologia da Informação**. 6. ed. São Paulo: Futura, 1998.

_____; KIM, J. **Keeping up with the quants**. Harvard Business Review Press, 2013.

_____, T. H. **Big Data at work:** dispelling the myths, uncovering the opportunities. Harvard Business Review Press, 2014.

DAVID, E. **Dados Abertos Governamentais**. 2009. Disponível em: <http://www.w3c.br/divulgacao/pdf/dados-abertos-governamentais.pdf>. Acesso em: 10 jul. 2015.

DEAN, J.; SANJAY, G. MapReduce: simplified data processing on large clusters. **Communications of the ACM**, v. 51, n. 1, p. 107-113, 2008. Disponível em: <http://static.googleusercontent.com/media/research.google.com/pt-BR//archive/mapreduce-osdi04.pdf>. Acesso em: 10 jan. 2016.

DEMCHENKO, Y.; GRUENGARD, E.; KLOUS, S. Instructional model for building effective Big Data curricula for online and campus education. In: 1ST IEEE STC CC AND RDA WORKSHOP ON CURRICULA AND TEACHING METHODS IN CLOUD COMPUTING, BIG DATA, AND DATA SCIENCE, Singapore, dez. 2014. Disponível em: <https://www.researchgate.net/publication/273945502_Instructional_Model_for_Building_Effective_Big_Data_Curricula_for_Online_and_Campus_Education>. Acesso em: 10 jan. 2016.

_____; LAAT, C; MEMBREY, P. Defining architecture components of the Big Data Ecosystem. In: 2 ND BDDAC2014 SYMPOSIUM, CTS2014 CONFERENCE, Minneapolis, maio 2014. Disponível em: <http://www.uazone.org/demch/presentations/bddac2014-bigdata-architecture-v01.pdf>. Acesso em: 10 jan. 2016.

DUBLIN CORE METADATA INITIATIVE – DMCI. **Dublin core metadata element set**, version 1.1. 2012. Disponível em: < http://dublincore.org/documents/dces/>. Acesso em: 15 jul. 2015.

GANTZ, J.; REINSEL, D. Extracting value from chaos. **IDC Iview**, v. 1142, n. 2011, p. 1-12, 2011. Disponível em: <http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf> Acesso em: 23 jun. 2015.

GOMES, H. E. **Manual de elaboração de tesauros monolíngues**. Brasília: Programa Nacional de Bibliotecas de Ensino Superior, 1990.

GRUBER, T. R. **What is an ontology?** 1993. Disponível em: <http://www-ksl.stanford.edu/kst/what-is-an-ontology.html> Acesso em: 10 dez. 2015.

GUARINO, N. Formal ontology and information systems. In: PROCEEDINGS OF FOIS'98, Trento, Italy, p. 81-97, jun. 1998. Disponível em: <http://www.mif.vu.lt/~donatas /Vadovavimas/Temos/OntologiskaiTeisingas KoncepcinisModeliavimas/papildoma/Guarino98-Formal%20O ntology%20and%20Information%20Systems.pdf>. Acesso em: 10 jan. 2016.

INDRAWAN-SANTIAGO, M. Database research: are we at a crossroad? Reflection on NoSQL. In: FIFTEENTH INTERNATIONAL CONFERENCE ON NETWORK-BASED INFORMATION SYSTEMS, p. 45-51, 2012.

ISOTANI, S.; BITTENCOURT, I. I. **Dados Abertos Conectados:** Em busca da Web do Conhecimento. São Paulo: Novatec, 2015.

KIMBALL, R., ROSS, M. **The Data Warehouse toolkit**: the definitive guide to dimensional modeling. 3. ed. Indiana: John Wiley & Sons, 2013.

LANEY, D. **Application delivery strategies.** Meta Group. 2001. Disponível em: <http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf> Acesso em: 23 jan. 2016.

MACEDO, F. L. O. **Arquitetura da Informação:** aspectos epistemológicos, científicos e práticos. 2005. Dissertação (Mestrado) – Departamento de Ciência da Informação e Documentação, Universidade de Brasília, Brasília.

MAUAD, T; *et al.* Análise comparativa entre distritos industriais: uma aplicação do enfoque sistêmico para avaliar projetos de desenvolvimento local. In: PROCEEDINGS OF THE THIRD INTERNATIONAL CONFERENCE OF THE IBEROAMERICAN ACADEMY OF MANAGEMENT. 2003. Disponível em: <http://www.fgvsp.br/iberoamerican/Papers/0112_Artigo%20IAM_final%20formatado.pdf> Acesso em: 4 dez. 2015.

MORROGH, E. **Information architecture:** An emerging 21st century profession. Pearson Education, 2002.

OPEN GOVERNMENT PARTNERSHIP – OGP. **Open by Default, Policy by the People, Accountability for Results**, 2011. Disponível em: <http://www.opengovpartnership.org/sites/default/files/091116_OGP_Booklet_digital.pdf> Acesso em: 10 jul. 2015.

ROSENFELD, L.; MORVILLE, P. **Information architecture for the world wide web**. California: O'Reilly Media, Inc., 2002.

SCHMIDT, E. Every 2 days we create as much information as we did up to 2003. **TechCrunch**, 2010. Disponível em: <http://techcrunch.com/2010/08/04/schmidt-data/>. Acesso em: 23 jan. 2016.

SHEARER, C. The CRISP-DM model: the new blueprint for Data Mining. **Journal of Data Warehousing**, v. 5, n. 4, p. 13-22, 2000. Disponível em: <https://mineracaodedados.files.wordpress.com/2012/04/the-crisp-dm-model-the-new-blueprint-for-data-mining-shearer-colin.pdf>. Acesso em: 10 jan. 2016.

SCHIESSL, M. **Lexicalização de ontologias:** o relacionamento entre conteúdo e significado no contexto da recuperação da informação. 2015. Tese (Doutorado) – Programa de Pós-Graduação em Ciência da Informação, Faculdade de Ciência da Informação, Universidade de Brasília, Brasília. Disponível em: <http://repositorio.unb.br/bitstream/10482/18663/1/2015_MarceloSchiessl.pdf>. Acesso em: 10 jun. 2016.

SHIN, D. H.; CHOI, M. J. Ecological views of Big Data: perspectives and issues. **Telematics and Informatics**, v. 32, n. 2, p. 311-320, maio 2015.

STUDER, R.; BENJAMINS, R. R.; FENSEL, D. Knowledge engineering: principles and methods. **Data & Knowledge Engineering**, v. 25, n. 1-2, p. 161-197, 1998. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0169023X97000566>. Acesso em: 10 jan. 2016.

TRIBUNAL DE CONTAS DA UNIÃO – TCU. Secretaria de Fiscalização de Tecnologia da Informação. **5 motivos para a abertura de dados na Administração Pública**. Brasília, 2015. Disponível em: <http://portal3.tcu.gov.br/portal/pls/portal/docs/2689107.PDF>. Acesso em: 23 jun. 2015.

USCHOLD, M.; GRÜNINGER, M. Ontologies: principles, methods and application. **The Knowledge Engineering Review**, v. 11, n. 2, p. 93-136, 1996. Disponível em: <http://www.upv.es/sma/teoria/sma/onto/96-ker-intro-ontologies.pdf>. Acesso em: 10 jan. 2016.

VICTORINO, M. C. **Organização da informação para dar suporte à arquitetura orientada a serviços:** reuso da informação nas organizações. 2011. Tese (Doutorado) – Programa de Pós-Graduação em Ciência da Informação, Faculdade de Ciência da Informação, Universidade de Brasília, Brasília. Disponível em: <http://repositorio.unb.br/bitstream/10482/10056/1/2011_MarcioCarvalhoVictorino.pdf>. Acesso em: 10 jan. 2016.

WODTKE, C.; GOVELLA, A. **Information architecture:** Blueprints for the Web. Pearson Education India, 2011.

WURMAN, R. S. **Ansiedade de informação 2:** um guia para quem comunica e dá instruções. São Paulo: Cultura, 2005.