Instituto Serzedello Corrêa

O uso de ferramentas de *Data Analytics* pelo Auditor Governamental

Leonardo Marques Garcia

Orientador: Alessandro de Oliveira Borges

Coletânea de Pós-Graduação, v.1 n.7

Auditoria Financeira

TRIBUNAL DE CONTAS DA UNIÃO

REPÚBLICA FEDERATIVA DO BRASIL

TRIBUNAL DE CONTAS DA UNIÃO

MINISTROS

José Mucio Monteiro (Presidente)
Ana Arraes (Vice-presidente)
Walton Alencar Rodrigues
Benjamin Zymler
Augusto Nardes
Aroldo Cedraz de Oliveira
Raimundo Carreiro
Bruno Dantas
Vital do Rêgo

MINISTROS-SUBSTITUTOS

Augusto Sherman Cavalcanti
Marcos Bemquerer Costa
André Luís de Carvalho
Weder de Oliveira

MINISTÉRIO PÚBLICO JUNTO AO TCU

Cristina Machado da Costa e Silva (Procuradora-Geral)
Lucas Rocha Furtado (Subprocurador-geral)
Paulo Soares Bugarin (Subprocurador-geral)
Marinus Eduardo de Vries Marsico (Procurador)
Júlio Marcelo de Oliveira (Procurador)
Sérgio Ricardo Costa Caribé (Procurador)
Rodrigo Medeiros de Lima (Procurador)

Instituto Serzedello Corrêa
Escola Superior do Tribunal de Contas da União

PROJETO GRÁFICO E CAPA
Núcleo de Comunicação - NCOM/ISC

Instituto Serzedello Corrêa
Escola Superior do Tribunal de Contas da União

PÓS-GRADUAÇÃO EM AUDITORIA FINANCEIRA

O uso de ferramentas de *Data Analytics* pelo Auditor Governamental

Leonardo Marques Garcia

Orientador(a):
Alessandro de Oliveira Borges, Especialista

Brasília | 2019

Resumo

O crescimento exponencial no volume e na variedade dos dados produzidos pelas diversas atividades da sociedade se apresenta como um grande desafio para as organizações detentoras destes dados. Todavia, as ferramentas para análise desse grande volume de dados, o chamado Big Data, evoluíram notavelmente nos últimos anos, auxiliando a avaliação de resultados e a tomada de decisão de gestores privados e públicos.

No setor público, o uso de tecnologias de análise de dados sobre o Big Data governamental pode ajudar a controlar adequadamente os gastos públicos, o que é uma atribuição da auditoria governamental.

Nesse contexto, o trabalho se propôs a avaliar como o uso de ferramentas e algoritmos baseados em grafos para Análise de Dados, como o Neo4j e a Cypher, podem ajudar o exercício do auditor governamental.

Palavras-chave: Big Data, Data Analytics, Setor Público, Auditor Governamental, Neo4j, Cypher.

Abstract

The exponential growth in the volume and variety of data produced by the various activities of society became a major challenge for organizations holding these data. However, the tools for analyzing this large volume of data, the so-called Big Data, having evolved remarkably in the last years, helping in the evaluation of results and decision making of managers and public administrators.

In the public sector, the use of data analysis technologies on the government's Big Data may help the proper control of public expenditures, which is an assignment of government auditing.

In this context, the paper proposes to evaluate how the use of graph algorithms and graph analytics tools for data analysis, as Neo4J and Cypher, can help the exercise of the government auditor.

Keywords: Big Data, Data Analytics, Public Sector, Government Auditor, Neo4j, Cypher.

Lista de Figuras

Sumário

1. Introdução

Nos dias atuais, os dados são produzidos em maiores quantidades e pelas mais variadas fontes. No entanto, a capacidade e a sofisticação das ferramentas para análise desse grande volume de dados, o *Big Data*, têm-se desenvolvido quase no mesmo ritmo. Nesse contexto, o alcance e a aplicabilidade do contexto do *Big Data* parece ilimitado (TOMAR, 2016).

Este incremento do volume de dados produzidos e armazenados não é observado em um setor específico. Seja no mercado privado ou no setor público, os desafios em relação à necessidade de se extrair valor das fontes de dados estão presentes e demandando soluções complexas e efetivas.

No que diz respeito ao setor público, frente às atuais restrições fiscais, o emprego de tecnologias inovadoras, porém acessíveis do ponto de vista financeiro, tais como a utilização de ferramentas de Análise de Dados sobre o *Big Data* governamental, pode se apresentar como um mecanismo eficiente para garantir a boa gestão dos recursos públicos.

Dentre as áreas para aplicação de tais ferramentas, destaca-se o campo de atuação dos órgãos de controle, que fazem uso das mais variadas fontes de dados para planejar e executar o seu escopo de atuação. Nesse contexto, o presente trabalho aborda a aplicação de ferramentas de *Data Analytics* no contexto de execução das auditorias governamentais.

1.1 Objetivo

O objetivo geral do trabalho consiste em avaliar de que forma o emprego de técnicas de Análise de Dados ou *Data Analytics* podem potencializar o trabalho do auditor público a partir da ampliação da visão deste profissional sobre o objeto auditado, para que possa, a partir de tais ferramentas, detectar e corrigir tempestivamente possíveis deficiências, erros e/ou fraudes e, consequentemente, assegurar a eficiente e eficaz gestão dos recursos públicos e a efetividade na implementação de políticas públicas.

Ao fim do trabalho espera-se que este contribua para a disseminação da cultura de aplicação de técnicas de *Data Analytics* no âmbito da Auditoria do Setor Público, de forma que novas ferramentas e/ou tecnologias específicas possam ser incorporadas aos processos de trabalho dos responsáveis pelo controle e fiscalização dos recursos públicos.

1.2 Estrutura do trabalho

O presente trabalho apresenta a seguinte divisão de capítulos:

- **Capítulo 2:** Aborda uma contextualização histórica e teórica a respeito do conceito de *Big Data,* a evolução dos números que o envolvem, bem como suas possíveis aplicações, com ênfase no setor público;

- **Capítulo 3:** Apresenta uma argumentação quanto à utilização de ferramentas de Análise de Dados – *Data Analytics*, no contexto do *Big Data*, com uma abordagem direcionada para a aplicação destas tecnologias no âmbito das auditorias governamentais;

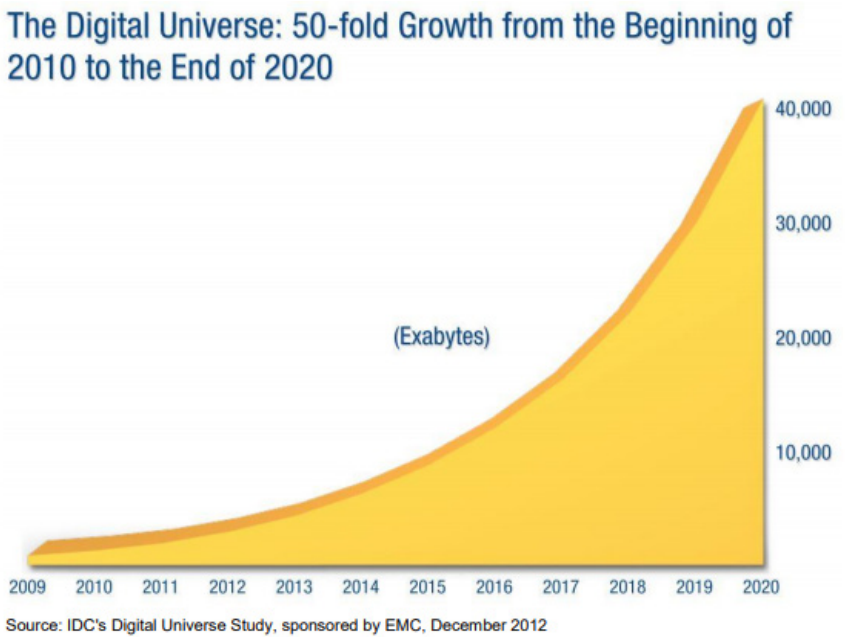
- **Capítulo 4:** Conduz um experimento introdutório, no qual a ferramenta Neo4j é explorada a fim de se avaliar a possibilidade do seu uso como instrumento de *Data Analytics* no contexto de uma auditoria governamental.

2. A era do *Big Data*

Segundo (IBM, 2013), a sociedade moderna globalizada gera, diariamente, quintilhões de bytes de dados. Conforme o estudo "*A Universe of Opportunities and Challenges*", desenvolvido pela consultoria EMC (GANTZ, 2012), o volume de bytes gerados anualmente entre 2006 e 2010 cresceu de 166 Exabytes para 988 Exabytes.

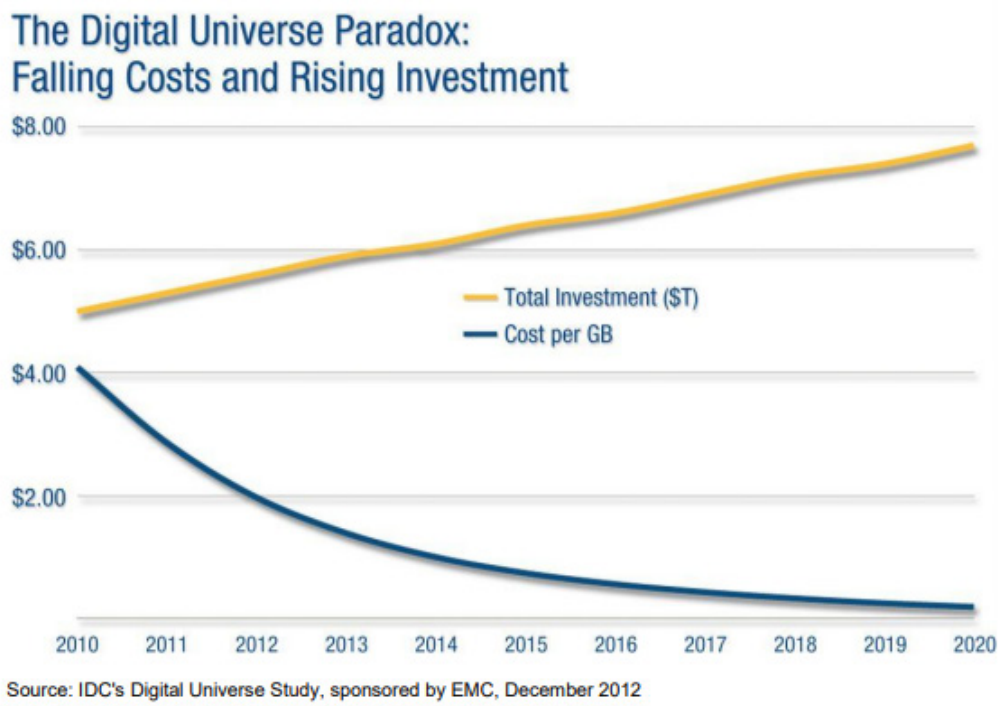
De acordo com a Figura 1, extraída daquele estudo, a perspectiva é que o volume de dados alcance, em 2020, a casa dos 40.000 Exabytes (40 vezes 10 elevado à 22ª potência).

Figura 1– Estimativa de crescimento do volume de dados digitais de 2010 a 2020



No mesmo estudo, a previsão foi que o custo do investimento por gigabyte entre 2012 e 2020 cairia de U$ 2,00 para U$ 0,20, conforme a figura 2, a seguir.

Figura 2 – Perspectiva de redução do custo por GB de dados entre 2010 e 2020

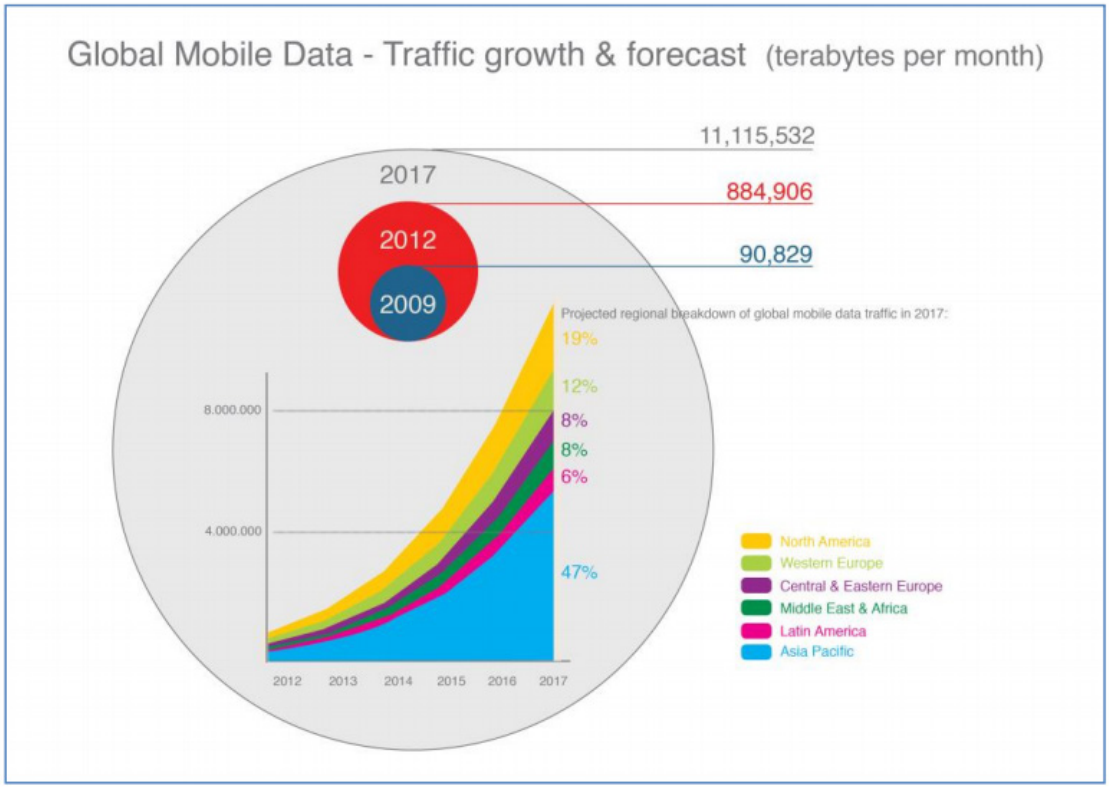


Contribuem para este crescimento exponencial de volume de dados as mais diversas ações diárias da sociedade, sejam em nível pessoal ou profissional; os registros corporativos e as movimentações financeiras das organizações; além de dispositivos e equipamentos que também geram dados.

O crescimento do número de smartphones, assim como dos usuários móveis, do tráfego de vídeo, da velocidade das redes 4G e 5G e da Internet das Coisas de forma exponencial incrementará este volume de dados gerados, em grande parte, em decorrência da explosão do tráfego de dados em dispositivos móveis, conforme representa a Figura 3 (LETOUZÉ, 2014).

Dessa forma, extrair inteligência e valor desse volume de dados variado e em crescimento exponencial amplia dos desafios no campo do *Big Data*. Neste contexto, a geração de volumes extraordinários de dados por organizações, pessoas e equipamentos é que subsidiou a nomeação da chamada Era do *Big Data*.

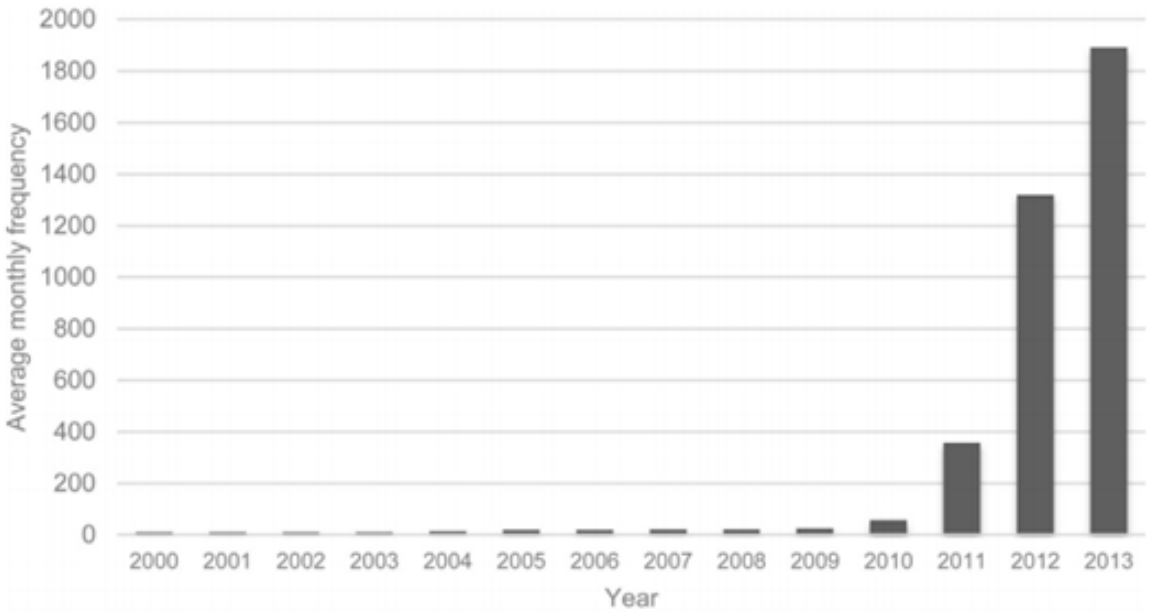
Figura 3 – Crescimento do Tráfego de Dados Móveis entre 2012 e 2017 (LETOUZÉ, 2014).



São muitas as versões sobre a origem do conceito de *Big Data*. Uma das mais aceitas remete ao início dos anos 90, quando a NASA começou a utilizar o termo *Big Data* para se referir aos complexos e volumosos conjuntos de dados que desafiavam os limites de processamento e armazenamento da computação convencional do período.

Apesar das referências a meados dos anos 90, a Figura 4 (GANDOMI, 2015) mostra que o termo se difundiu a partir de 2011. Segundo (GANDOMI, 2015), o atual entusiasmo pode ser atribuído às iniciativas promocionais das empresas de tecnologia líderes do mercado que investiram na construção de um mercado de análise a partir daquele período.

Figura 4 - Frequência do termo "Big Data" na biblioteca de pesquisa da ProQuest



Big Data é um termo que descreve o grande volume de dados gerados por pessoas, empresas, equipamentos, etc., nos mais variados formatos.

Este termo também pode ser definido como uma coleção de bases de dados tão variada e volumosa que a execução de operações relativamente simples sobre elas, tais como captura, armazenamento, gestão, análise, remoção, ordenação e/ou sumarização dos dados, se tornam muito complexas.

Segundo (KAUFMAN, 2013), o termo é uma combinação de tecnologias de gestão de dados que evoluíram ao longo dos anos. Permite que as organizações armazenem, administrem e manipulem grandes quantidades de dados na velocidade e tempo corretos para conseguir os conhecimentos certos.

Como a maioria das organizações está na fase inicial de suas jornadas com *Big Data*, muitas delas ainda estão experimentando técnicas que as permitem coletar quantidades massivas de dados para determinar se existem padrões escondidos, por exemplo. Implementar tais soluções requer uma infraestrutura tecnológica apropriada para apoiar a escalabilidade, a distribuição e a administração desses dados.

Contudo, o fator mais importante não é a quantidade de dados disponíveis na estrutura implementada, mas sim o que é feito com eles. No contexto de uma organização, seja ela pública ou privada, o foco principal da implementação de uma solução de *Big Data* reside

em analisar os dados a fim de se obter insights que possam aprimorar o processo de tomada de decisões e a definição de ações estratégicas para o negócio da organização.

Dessa forma, para (KAUFMAN, 2013), a chave para entender o termo *Big Data* é compreender que os dados devem ser administrados para que possam alcançar as necessidades de negócio que uma dada solução é projetada para resolver.

2.1 As dimensões do *Big Data*

No início dos anos 2000, o conceito de *Big Data* ganhou visibilidade quando Doug Laney (LANEY, 2001) deu uma definição para o termo com os chamados três "V": Volume, Velocidade e Variedade.

Segundo a SAS (SAS, 2019), uma das principais empresas fornecedoras de soluções para *Big Data* em nível global, a definição para os três "V" seria a seguinte:

- **Volume:** Organizações coletam dados de fontes variadas, incluindo transações financeiras, mídias sociais e informações de sensores ou dados transmitidos de máquina para máquina. No passado, armazená-los teria sido um problema, mas o desenvolvimento de novas tecnologias tem facilitado o processo de armazenamento do *Big Data.*

- **Velocidade:** Os dados são transmitidos em uma velocidade sem precedentes e devem ser tratados em tempo hábil. Etiquetas RFID[1], sensores e medições inteligentes estão impulsionando a necessidade de lidar com fluxos de dados praticamente em tempo real.

- **Variedade:** Os dados são gerados em inúmeros formatos — desde estruturados (informação organizada em campos alfanuméricos, em bases de dados tradicionais) a não-estruturados (documentos de texto, e-mail, vídeo, áudio, etc.).

Todavia, o termo não é imutável e ao longo dos últimos anos novas dimensões foram adicionadas à definição do termo *Big Data*. Dentre essas novas dimensões, a destacam-se as seguintes (SAS, 2019):

- **Variabilidade:** Além das crescentes velocidade e variedade dos dados, seus fluxos podem ser altamente inconsistentes com picos periódicos.

1 Etiquetas RFID (Radio Frequency Identification), ou etiquetas de Identificação por Rádio Frequência, são dispositivos (chips) de identificação e rastreamento por meio de um pequeno sinal de radiofrequência.

- **Complexidade:** Os dados de hoje vêm de múltiplas fontes, o que torna difícil ligá-los, combiná-los, limpá-los e transformá-los entre sistemas. No entanto, é necessário conectar e correlacionar relações, hierarquias e ligações múltiplas, a fim de evitar perder o controle sobre seus dados.

Por fim, mais recentemente foram adicionados mais dois "V" à definição cunhada pelo analista Doug Laney (MARR, 2014):

- **Veracidade:** Para colher bons frutos do processo do *Big Data* é necessário obter dados verídicos, de acordo com a realidade. Um dos pontos mais importantes de qualquer informação é que ela seja verdadeira. Por exemplo, com o *Big Data* não é possível controlar cada hashtag do *Twitter* ou notícia falsa na internet, mas com análises e estatísticas de grandes volumes de dados é possível compensar as informações incorretas, de forma que se obtenha dados que agreguem valor ao processo.

- **Valor:** Ter acesso a uma grande quantidade de informação a cada segundo não adianta nada se não gerar valor para o negócio, seja ele qual for. É sempre importante lembrar dos custos e benefícios do processo de análise, a fim de se buscar agregar valor ao que se está fazendo. O valor agregado pelo processo de *Big Data* desenvolvido, seja com coleta, armazenamento e análise de dados, tem que compensar os custos financeiros envolvidos.

As dimensões anteriormente definidas identificam os principais desafios que precisam de ser avaliados em um ambiente de *Big Data*. No entanto, antes de se planejar o armazenamento dos dados em um *Big Data*, é fundamental identificar todas as potenciais fontes de dados, considerando as decisões que serão tomadas ao se fazer uso destes dados, destacando as seguintes: Como armazenar e gerir os dados? Quanto analisar? O que fazer com as análises?

No passado recente, os custos do armazenamento e gerenciamento dos dados seriam proibitivos para as pequenas e médias empresas e para a maioria das organizações públicas, frente às priorizações orçamentárias. Contudo, nos dias atuais já existem opções de mais baixo custo para se implementar soluções de *Big Data*. Essa é uma etapa em que a ênfase seria nas dimensões Volume, Velocidade e Variedade.

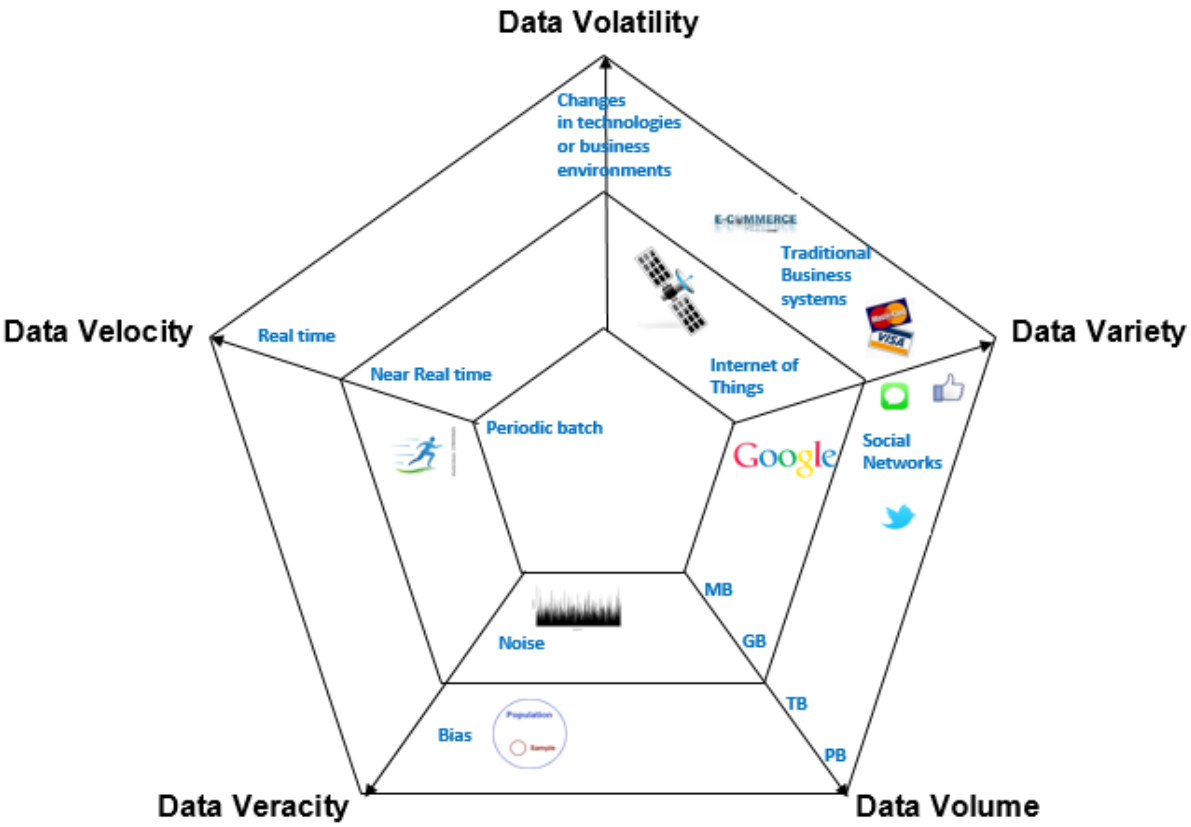
A despeito da redução dos custos, é importante determinar previamente quais dados são relevantes, antes de armazená-los e, consequentemente, analisá-los. Algumas organizações com mais recursos não excluem quaisquer dados de suas análises, mas para boa parte delas é fundamental a seleção de parte dos dados. No entanto, a análise não resta prejudicada, já que as próprias soluções de *Big Data*, por meio de processos estatísticos, permitem análises com excelente acurácia, ainda que esta não

tenha sido feita sobre toda a população dos dados. Como se percebe, nessa etapa a ênfase se dá na dimensão Veracidade.

Por fim, o mais importante é a forma de se utilizar os insights descobertos, ou seja, a dimensão Valor se destaca, já que para a organização, quanto maior o conhecimento sobre a sua atividade fim, melhores decisões de negócio serão tomadas.

A despeito da lista de "Vs" ter crescido com o tempo, a figura (HAMMER, 2017) a seguir, baseada no modelo de Doug Laney, enfatiza as principais dimensões e, consequentemente, as oportunidades e desafios que as organizações enfrentam quando incorporam *Big Data* em suas atividades fim.

Figura 5 – Dimensões do Big Data



2.2 O *Big Data* em números

A Forbes Insights (PRESS, 2015), patrocinada pela empresa Teradata em parceria com a consultoria McKinsey, entrevistou 316 executivos de grandes empresas globais a

fim de fornecer uma visão do estado das implementações de análise de *Big Data* em nível mundial. Os principais destaques da pesquisa são a seguir apresentados:

- 90% dos entrevistados relataram níveis médios a altos de investimento em análise de *Big Data*, e cerca de um terço chamou seus investimentos de "muito significativos";

- Dois terços relataram que as iniciativas de *Big Data* tiveram um impacto significativo nas receitas;

- 59% consideram *Big Data* uma das cinco principais questões ou a única maneira mais importante de obter uma vantagem competitiva;

- 51% dos executivos disseram que adaptar e refinar uma estratégia baseada em dados é a maior barreira cultural e 47% relataram colocar o aprendizado de *Big Data* em ação como um desafio operacional;

- 43% citaram a promoção de uma cultura que premia o uso de dados e valoriza a criatividade e a experimentação com dados como principais desafios;

- 51% das organizações em que o *Big Data* é visto como a maneira mais importante de obter vantagem competitiva são lideradas por CEOs que se concentram pessoalmente em iniciativas de *Big Data*;

- Em organizações em que o *Big Data* é visto como uma das cinco principais questões que recebe tempo e atenção significativos do topo da administração, o patrocinador normalmente está um nível abaixo do líder principal;

- 48% dos executivos entrevistados consideram as decisões de negócios baseadas em dados como um desafio estratégico;

- 43% citam o desenvolvimento de uma estratégia corporativa como um obstáculo significativo;

- 46% dos executivos entrevistados relataram que a contratação de talentos capacitados na área de dados é um desafio.

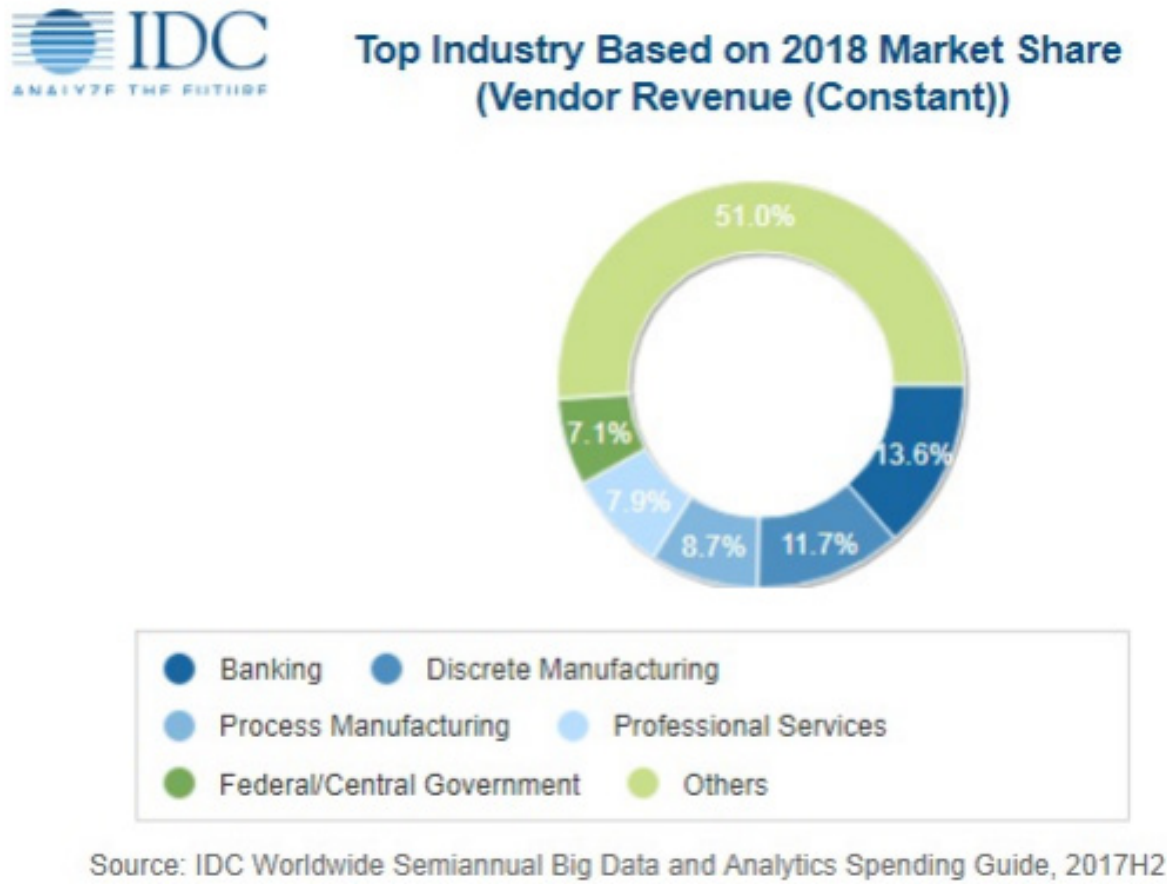
Nos anos seguintes, outras publicações passaram a apresentar a evolução do mercado de *Big Data*, seja no nível de receitas ou nos investimentos, corroborando os destaques da Forbes Insigths.

O *International Data Corporation* – IDC publicou em outubro de 2018 o seu mais recente *Worldwide Semiannual Big Data and Analytics Spending Guide* (IDC, 2018), prevendo

que a receita mundial para soluções de *Big Data* e *Analytics* alcançará U$ 260 bilhões em 2022, com uma taxa de crescimento anual composta de 11,9% sobre a previsão 2017-2022 período.

Os setores previstos para fazerem os maiores investimentos em soluções de *Big Data* são os setores bancários, a indústria, o varejo, o setor de serviços e o setor governamental. Combinados, esses setores serão responsáveis por quase metade (U$ 81 bilhões) das receitas mundiais com *Big Data* e *Analytics*. Eles também serão os setores com maior investimento em 2022, quando o investimento somado de todos será de U$ 129 bilhões.

Figura 6 – Participação dos Setores Econômicos na Receita com *Big Data* (IDC, 2018)



O relatório da IDC encontrou inúmeras razões pelas quais certos setores estão investindo em *Big Data* e análise de dados dos negócios. No setor bancário, os investimentos se concentrarão principalmente em questões relacionadas à segurança e conformidade. Para os outros setores, há um foco em obter mais informações do cliente, melhorando a experiência destes clientes e também visando o incremento das vendas.

No período sob análise, os Estados Unidos são o maior mercado, entregando quase U$ 88 bilhões receitas em 2018 e mais da metade do total mundial ao longo da previsão de cinco anos. A Europa Ocidental é o segundo maior mercado, com expectativa de receitas de U$ 35 bilhões, seguido pela região Ásia/Pacífico, com U$ 23,9 bilhões de receitas esperadas. O Japão será o segundo maior país para investimentos em 2018, seguido pelo Reino Unido, Alemanha e China.

No que diz respeito ao mercado latino americano, os dados mais recentes se referem aos números destacados no relatório "*Latin American Big Data and Analytics (BDA) market*" (FROST, 2018). Segundo o documento, o mercado brasileiro de *Big Data* e *Analytics* (BDA) movimentou em 2017 uma receita total de U$ 1,35 bilhão. O estudo aponta que o Brasil representa 46,7% do mercado latino-americano, que movimentou em 2017 receitas no valor de U$ 2,9 bilhões, apresentando uma taxa de crescimento composto anual de 19,2% até 2023, quando este mercado deve chegar a U$ 8,5 bilhões.

2.3 Aplicações para o *Big Data*

O artigo intitulado "*Big Data: The Next Frontier For Innovation, Competition And Productivity*" (CHUI, 2011) apresentou algumas áreas com maior potencial para aplicações de soluções de *Big Data*.

No âmbito das organizações privadas, a fim de se obter vantagens competitivas, destacou-se que o armazenamento dos dados transacionais em formato digital permite uma análise que visa obter informações precisas e detalhadas sobre o negócio, como exemplo: equilíbrio dos estoques com as estimativas de venda dos próximos meses, a fim de aprimorar os fluxos de caixa. Em outra medida, o *Big Data* contribui para o aprimoramento da relação com os clientes, permitindo uma segmentação focada no perfil destes clientes, seja no que diz respeito ao consumo de produtos ou ao uso de serviços. Além disso, as soluções visam também incrementar a produtividade da própria organização.

Adicionalmente, destaca-se que o potencial para aproveitamento dos grandes volumes de dados no setor público é enorme, tais como aplicações na área de saúde a fim de gerar gastos com maior eficiência e qualidade, reduzindo erros e fraudes. O *Big Data* pode ajudar a descobrir um valor significativo nestas bases de dados mediante a geração de informação transparente e utilizável em maior frequência (CHUI, 2011).

A chamada *Digital Health* diz respeito à utilização de soluções de *Big Data* para coletar, agregar, trabalhar e analisar dados estruturados e não estruturados do setor de saúde, a fim de gerar informações clínicas que proporcionem diagnósticos mais exatos, por meio do cruzamento de uma multiplicidade de variáveis clínicas.

Ainda no âmbito governamental, podem ser destacados outros benefícios do uso de soluções de *Big Data* (CHILE, 2013):

- Melhorar a eficiência da gestão pública e a qualidade das políticas públicas;
- Melhorar a qualidade dos dados por meio de uma maior transparência;
- Aprimorar a participação da cidadania;
- Agregar valor às informações e decisões governamentais;
- Fomentar a inovação por meio da utilização de dados abertos;
- Contribuir para o desenvolvimento de aplicações e serviços inovadores;
- Promover o crescimento econômico.

Como as decisões da Administração Pública envolvem difíceis escolhas, haja vista as limitações orçamentárias e a reserva do possível na aplicação dos recursos, o *Big Data* pode se apresentar como uma ferramenta eficiente para a priorização de programas e políticas públicas; para a prevenção de desastres naturais e epidemias; e para a otimização dos investimentos em infraestrutura.

Além disso, as soluções tecnológicas podem contribuir para que a Administração Pública combata a corrupção e o desvio de receitas, por exemplo: coletando, processando e cruzando informações de contribuintes a fim de combater a lavagem de dinheiro e outros crimes financeiros.

Uma aplicação real, que exemplifica o uso de ferramentas de *Data Analytics* para enfrentamento da corrupção no âmbito governamental, é o sistema ALICE, acrônimo de "análise de licitações e editais". A partir de consultas ao Diário Oficial da União e ao Comprasnet, o portal federal de aquisições, o ALICE coleta arquivos e dados de todas as licitações e de todas as atas de pregão publicadas. A partir do texto do edital, ele faz a obtenção do valor estimado da licitação e define o risco da licitação, após análises de texto, com foco em restrição de competitividade na habilitação (CGU, 2015).

Ainda, com o uso de *Big Data* é possível que as Administrações desenvolvam um sistema de monitoramento real time para que a população possa acompanhar o consumo de energia e as possibilidades de sobrecarga no fornecimento, incrementando a cidadania via transparência de dados que atualmente não são públicos.

Inclusive, indicadores de satisfação a respeito da atuação da Administração Pública podem ser criados a partir do monitoramento das mídias sociais por meio de ferra-

mentas de *Social Big Data*. Assim, uma Administração Pública moderna e orientada a dados pode obter insights para implantar novos projetos ou políticas para problemas detectados nestas mídias, e que possivelmente não chegariam ao conhecimento do poder público pelos meios tradicionais de comunicação.

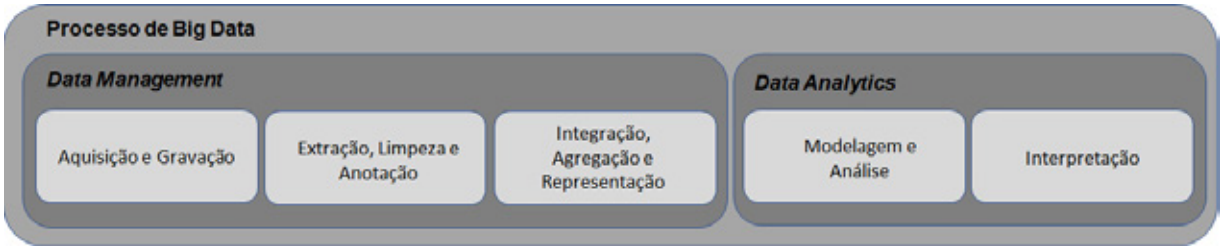
Finalmente, pode-se destacar um dos desdobramentos do termo *Big Data*, que é o conceito de *Data Analytics*, que se refere às ferramentas propriamente ditas, aplicadas na análise das bases de dados massivas, quais sejam: hardwares e/ou softwares capazes de tratar os dados, a fim de transformá-los em informações úteis às organizações.

As ferramentas de *Data Analytics* permitem analisar dados estruturados e/ou não estruturados, separadamente ou em conjunto, permitindo, tecnologicamente, a descoberta de informações em tempo real, enquanto ainda estavam fora do alcance humano.

A importância fundamental do *Data Analytics* reside no fato de que o *Big Data* é inútil caso esteja somente armazenado, sem qualquer processo de inteligência sobre as bases de dados. Seu valor potencial só é desbloqueado quando utilizado para impulsionar a tomada de decisões.

Para permitir essa tomada de decisão com base em evidências, as organizações, sejam públicas ou privadas, precisam de processos eficientes para transformar grandes volumes de dados, dinâmicos e em movimento rápido, em insights significativos. O processo geral de extrair percepções do *Big Data* pode ser dividido em cinco etapas, conforme apresentado na Figura 7 (GANDOMI, 2015). Esses cinco estágios formam os dois subprocessos principais: *Data Management* (Gerenciamento de Dados) e *Data Analytics* (Análise de Dados).

Figura 7 - Processos para extrair insights de Big Data



O gerenciamento de dados envolve processos e tecnologias de suporte para: adquirir e armazenar dados; extrair, preparar e tratar dados; e integração, agregação e representação para análise.

Já o subprocesso *Data Analytics* se refere às técnicas usadas para modelar, analisar e adquirir inteligência em relação aos dados, sendo vista como o principal subprocesso no processo geral de extração de insights do *Big Data*.

3. *Big Data* como matéria-prima para o *Data Analytics*

Conforme apresentado anteriormente, a ampla disponibilidade de dados tem aumentado o interesse das organizações, privadas ou públicas, em métodos para extrair informações úteis e conhecimento a partir destes dados, o chamado *Big Data*. Nos últimos anos, as organizações têm ampliado os investimentos em infraestrutura de negócios a fim de melhorar a capacidade de coleta e análise do *Big Data* organizacional.

No passado, estas organizações contratavam equipes de estatísticos e analistas para explorar manualmente os conjuntos de dados. Contudo, o volume, a complexidade e a variedade dos dados superaram a capacidade da análise manual. Concomitantemente, as redes de comunicação, os hardwares e softwares se tornaram muito mais poderosos e foram desenvolvidos algoritmos que conectam conjuntos de dados para permitir análises muito mais amplas e profundas do que antes (PROVOST, 2013).

A convergência desses fenômenos deu origem à aplicação, cada vez mais difundida, das chamadas ferramentas de *Data Analytics*. Tais ferramentas são aplicadas no campo do *Big Data* a fim de explorar e analisar estes grandes volumes de dados em busca de padrões, previsões, erros, associações entre outros.

Com possibilidades de uso tanto no setor público quanto no privado, os gestores estão começando a entender que as tecnologias avançadas em *Big Data Analytics* permitirão adquirir insights mais elaborados, detalhados e valiosos, contribuindo para a tomada de decisão com informação qualitativa e mais tempestividade.

3.1 Ferramentas para o *Data Analytics*

O valor do dado digital bruto só pode ser mensurado quando este se transforma em informação e conhecimento, ou seja, quando é tratado e analisado. Neste contexto é que se inserem as ferramentas de *Data Analytics* ou “Análise de Dados”, que permitem uma análise sofisticada e complexa de um variado, volumoso e, possivelmente, não estruturado conjunto de dados, a fim de trazer os insights sobre o negócio ou atividade organizacional que nenhuma outra técnica seria capaz de produzir.

Dessa forma, pode-se dizer que o *Big Data* é a matéria-prima para as ferramentas de *Data Analytics,* que por sua vez se apresentam como novas formas de coleta, armazenamento e análise, contribuindo para a tomada de decisões em tempo real.

No mercado existem diversas soluções que contribuem para a tomada de decisões mais acertadas. No entanto, frente à gama de opções disponíveis, a escolha da ferramenta depende diretamente das demandas da atividade da organização e da capacitação dos recursos humanos, o que pode exigir ferramentas específicas.

Quanto às soluções comerciais, alguns dos principais fornecedores são as empresas: IBM, Microsoft, SAS e Oracle. O interessante a se destacar em relação às essas soluções comerciais é o fato de que elas foram desenvolvidas sendo suportadas por uma solução de *Software Open Source*, o *Apache Hadoop* (HADOOP, 2019).

Segundo a IBM (IBM, 2019), o *Apache Hadoop* oferece um processamento confiável, escalável e distribuído para grandes conjuntos de dados, usando modelos de programação simples e podendo ser construído em clusters de computadores comuns, o que fornece uma solução econômica para armazenamento e processamento de dados estruturados, semiestruturados e não estruturados, sem requisitos de formato.

O mapeamento de dados permite a priorização e atribuição de relevância. Quando combinado com ferramentas de análise, torna-se um poderoso instrumento para os usuários de negócios. Não por acaso, o *Apache Hadoop* está se tornando o núcleo da infraestrutura de *Big Data*, pois além de *open source* e gratuito, pode ser usado em *hardwares* de baixo custo, fator este fundamental para as organizações que objetivam reduzir custos de infraestrutura de TI e ainda capitalizar os benefícios do *Data Analytics.*

Em relação às soluções complementares em *Open Source*, destacam-se os Bancos de Dados NoSQL, como o MongoDB ou Neo4J, por exemplo. Bancos de Dados NoSQL são bancos de dados distribuídos e não-relacionais, projetados para tratar dados não-estruturados ou semi-estruturados, dados estes comuns nos tempos de *Big Data.* Dessa forma, tais ferramentas suprem a lacuna dos tradicionais Bancos de Dados relacionais, visto que estes estão limitados a tratar somente conjuntos de dados que podem ser armazenados em linhas e colunas, usando a linguagem SQL (*Structured Query Language*) para consultas e demais operações (definição, controle e transação).

Além dessas ferramentas, deve-se destacar as linguagens de programação mais populares no âmbito do *Data Analytics: Python e R* (KING, 2016).

A linguagem *Python*, que tem uma comunidade de desenvolvedores ativa e com muita documentação disponível no GitHub[2], apesar de ser uma linguagem de uso geral, devido aos seus módulos e pacotes para a área de *Data Science*, utilizá-la para gerar visualizações das informações garante que um grande volume de dados possa ser tratado e limpo (KAPPAL, 2017).

Já a linguagem estatística R, é conhecida por sua capacidade de processar estatísticas de grandes volumes de dados e criar gráficos sofisticados. Devido a este potencial, algumas organizações fornecedoras de ferramentas de *Data Analytics*, como Oracle (ORACLE, 2014) e Microsoft (MICROSOFT, 2018), a fim de aproveitarem um conjunto integrado de *software* para manipulação de dados, cálculo e exibição gráfica, passaram a adotar o R como linguagem padrão para análises estatísticas.

3.2 Data Analytics no Setor Público

Segundo (BONFIM, 2015), no Brasil, a Lei de Acesso à Informação (Lei federal nº 12.527/2011), institui-se não só como um marco jurídico-legal, mas também institucional, à sociedade brasileira e perante organismos internacionais para propiciar aos diversos atores da sociedade civil a primariedade, autenticidade, integridade e disponibilidade de consultar qualquer órgão público sobre informações de seu interesse, sem juízo de razão ou motivação, com o intuito de garantir condições sociais ao envolvimento democrático dos cidadãos brasileiros com organizações do Estado.

Assim, para atender uma demanda social e legal, os gestores públicos precisam disponibilizar informações com transparência e rapidez, sendo muito importante que as ações e políticas públicas de um governo sejam visíveis para toda a população, assim como os seus gastos e investimentos.

Todavia, nem sempre a disponibilização das informações é célere ou trivial, devido, entre outros fatores, aos diversos obstáculos que podem impedir que dados de diferentes fontes sejam visualizados em conjunto: existem dificuldades para a obtenção de algumas bases, problemas na própria confiabilidade dos dados, além de cobranças desnecessárias, entre os próprios órgãos, para disponibilização de informações que, em tese, deveriam ser públicas e/ou compartilhadas entre eles. Ainda que os dados estejam disponíveis, existem outros possíveis obstáculos para a coleta, tratamento e análise de dados no âmbito do setor público, em decorrência da existência de diferentes softwares e ferramentas sem interoperabilidade, o que prejudica a integração de todas as fontes de dados para fins de uma efetiva análise.

2 GitHub é uma plataforma de hospedagem de código-fonte que usa o sistema Git para controle de versão de projetos de software. Ela permite que diversas pessoas contribuam, simultaneamente, no mesmo projeto, editando e criando novos arquivos sem o risco de as versões serem sobrescritas.

Dessa forma, os gestores públicos podem ter dificuldades em analisar e exibir os seus próprios dados, a fim de atender as demandas de outros gestores ou dos cidadãos.

Assim, como o poder público age no interesse de todos, a despeito das possíveis dificuldades, incrementar a transparência das suas ações é fundamental para que exista uma relação de confiança entre sociedade e governo.

Uma das melhores formas de ser transparente e apresentar o desempenho da sua gestão pública é com dados confiáveis, por exemplo: quando for preciso explicar um investimento ou gasto de recurso público para a população, a clareza dos dados que o justificam será a melhor argumentação possível para o gestor.

No Brasil, a partir da publicação da Lei de Acesso à Informação, foram estabelecidas diversas iniciativas de transparência por meio da disponibilização de variadas informações sobre as ações do setor público. Então, ao aprimorar o processo de transparência, quando esta é baseada em análise de dados, verifica-se a melhoria do processo de tomada de decisões pelo poder público.

Frente à necessidade de aprimorar a sua gestão e garantir a transparência, os governos também devem implementar soluções de *Big Data Analytics* para melhorar a sua capacidade de servir os cidadãos e para enfrentar os grandes desafios públicos que envolvem a economia, saúde, emprego, segurança pública, meio ambiente, etc.

Como já destacado, os serviços governamentais são conhecidos pela criação de grandes volumes de dados, que em sua maior parte não estão integrados para análise em conjunto.

De um modo geral, *Data Analytics* sobre o grande volume de dados governamentais proporciona uma série de benefícios, dentre os quais podem ser citados:

- incremento da transparência da gestão;
- melhoria da tomada de decisão;
- melhora de resultados dos projetos governamentais;
- eliminação do desperdício e aumento da produtividade no serviço público;
- personalização da experiência dos cidadãos;
- redução da fraude nos impostos e na seguridade social;
- redução da criminalidade e das ameaças à segurança pública;

- diminuição do custo total na saúde.

Exemplificando com um caso brasileiro, o uso de *Data Analytics,* como suporte ao controle, pode ser empregado para ampliar o escopo das análises realizadas nos processos de aquisições governamentais a fim de compreender a estrutura de custos de um órgão ou entidade. Adicionalmente, soluções podem ser utilizadas a fim de mapear as empresas fornecedoras dos produtos, objetivando estabelecer preços de referência. Por conseguinte, o *Data Analytics* pode ser usado para identificar as possíveis redes de relacionamentos entre contratados e a Administração Pública (VAN ERVEN, 2015).

3.3 *Data Analytics* eo Auditor Governamental

No que diz respeito à Administração Pública, especificamente no campo do controle da atuação governamental, verifica-se que o processo de digitalização da gestão contábil, orçamentária, financeira, operacional e patrimonial dos órgãos e entidades, com a implementação de sistemas, tais como o SIAFI, SIASG, SIAPE e SICONV, etc., tem contribuído para subsidiar e aprimorar a auditoria governamental (COSTA, 2012).

Não somente o desenvolvimento dos sistemas de informação utilizados pelos órgãos e entidades contribuiu para fortalecer o controle da Administração Pública, mas também a própria evolução das tecnologias de informação e comunicação em nível mundial contribuiu para a introdução de novas ferramentas e metodologias de auditoria.

Nesse sentido, as ferramentas e/ou metodologias de Data Analytics se tornam aliadas do auditor público para que este faça uso de informações oportunas e confiáveis para que o seu órgão de controle possa expressar sua opinião adequadamente. Além disso, com esta atuação o auditor pode contribuir para que o gestor público melhore sua gestão por meio do aprimoramento do seu processo de tomada de decisão. Dessa forma, quando a auditoria governamental é eficiente e confiável, o processo de transparência pública se desenvolve.

Com o auxílio dessas ferramentas, um trabalho de auditoria que seria tedioso, manual, complexo e demorado pode ser automatizado. Além disso, a extensão e o momento da auditoria pode ser expandido para cobrir toda a população e volume de dados. Assim, é possível realizar análises mais confiáveis e abrangentes que permitem identificar e visualizar possíveis desvios que requerem uma análise mais aprofundada (KUENKAIKAEW, 2013).

Adicionalmente, com o uso de métodos analíticos, os auditores governamentais não só podem examinar eventos passados a fim de criar ajustes com base em alterações, erros ou fraudes que já ocorreram, mas também têm condições de realizar uma auditoria que busca detectar ou prevenir impropriedades e/ou irregularidades.

O uso de ferramentas de Análise de Dados permite transformar a auditoria de um órgão ou entidade em um processo contínuo, de forma que, por exemplo, padrões de gastos até então nunca observados sejam detectados e correlacionados de forma mais eficaz, permitindo uma análise mais célere e objetiva.

As possibilidades de armazenar, processar e cruzar dados para estabelecer vínculos, associações e correlações, facilitando o processo de obter inteligência a partir de sua capacidade de efetuar desagregações e análises multidimensionais é o principal diferencial dessas tecnologias de *Data Analytics.*

Neste contexto tecnológico, o principal desafio a ser enfrentado pelo auditor é conseguir utilizar de forma adequada e eficiente todas as técnicas e ferramentas que têm à sua disposição. Uma possível inexperiência na utilização destas ferramentas de *Data Analytics* contribui para uma interpretação inadequada dos dados, impedindo que o auditor alcance o objetivo do seu trabalho, qual seja: uma auditoria mais eficiente e consequentemente mais econômica (MARQUES, 2016). Portanto, é fundamental que os órgãos de controle não somente invistam em tais ferramentas, mas capacitem seus recursos humanos para que desenvolvam habilidades no uso delas.

No entanto, o auditor não pode ser visto como uma engrenagem isolada no contexto da auditoria com Análise de Dados. O *analytics* é mais efetivo quando apresenta uma abordagem integrada, de forma que estes auditores designados para o trabalho atuem com profissionais de ciência de dados e/ou tecnologia da informação, bem como com especialistas no objeto auditado, quando necessário (DELOITTE, 2016).

Figura 8 – Abordagem Integrada de Auditoria com Análise de Dados (DELOITTE, 2016)

Neste contexto, quando a equipe de auditoria estiver planejando e buscando áreas para serem auditadas, a partir de fontes de dados, é importante que se faça os questionamentos adequados para se definir o real escopo da auditoria. O chamado direcionamento inteligente à auditoria (DELOITTE, 2016), com o emprego de *Data Analytics*, relaciona os questionamentos iniciais da equipe com hipóteses analíticas a serem testadas após o processo de limpeza, tratamento e análise dos dados, para posterior relato e comunicação dos resultados, conforme ilustrado a seguir.

Figura 9 – Metodologia de direcionamento inteligente à auditoria (DELOITTE, 2016)

Tal direcionamento permite realizar uma auditoria mais ágil, focada e inovadora. No entanto, para tornar o *Data Analytics* possível no contexto da auditora governamental, é necessário contar com a adaptação de gestores e servidores, processos, normativos, dados, infraestrutura e tecnologia, a fim de incorporar este processo na estratégia do ente, a fim de, efetivamente, gerar valor por meio de sua aplicação.

4. Experimento com ferramenta de *Data Analytics*

Na atual era do *Big Data*, a informação de valor, gerada a partir dos dados, se tornou um dos principais ativos das organizações, sejam elas privadas ou públicas. Mais recentemente, a expressão criada pelo matemático londrino especializado em ciência de dados, Clive Humby, que afirma que os “dados são o novo petróleo” (MARR, 2018), ganhou as manchetes ao destacar o quão valioso são os dados produzidos.

Tal expressão é usada para defender a ideia de que os dados são tão valiosos quanto o petróleo, mas, assim como este, precisam ser “refinados”, isto é, analisados e/ou tratados. Nesse contexto, em tese, só tem a ganhar quem souber fazer bom uso para aproveitar todo o potencial dos dados.

Nesse sentido, a enormidade dos dados disponíveis vem incrementando o interesse das organizações em tecnologias que permitam extrair informações úteis e conhecimento a partir destes dados. Dessa forma, os profissionais e as organizações precisam, continuamente, serem alfabetizados em dados a fim de conseguir usá-los em seu potencial máximo.

Como já destacado, o dado bruto em si não é o ativo mais valioso, mas sim a inteligência por trás deles é o que determina seu real valor. Portanto, o uso de ferramentas de *Data Analytics* que permitem analisar o conjunto dos dados, apontando de maneira assertiva os melhores insights, pode transformar efetivamente a realidade de atuação das organizações.

Com possibilidades de uso também na administração pública, as tecnologias avançadas em *Data Analytics* podem contribuir para a obtenção de análises mais qualitativas, contribuindo para melhoria da tomada de decisão do gestor público, ou mesmo contribuindo para o controle da sua atuação.

Para que o emprego de *Data Analytics* se desenvolva no contexto do controle e/ou fiscalização da atuação governamental, é necessário contar com o patrocínio da administração, a fim de se aprimorar os processos operacionais e normativos dos entes, além de se investir em infraestrutura tecnológica e capacitação do corpo funcional, de forma que a implementação de uma metodologia contínua de análise de dados da administração pública possa agregar valor, não se comportando como uma nova despesa corrente.

Nesse contexto, e tendo como motivação a abordagem e metodologia de ensino apresentada pelo professor responsável pela disciplina “Análise de Dados”, da Especialização em Auditoria Financeira, buscou-se no presente trabalho apresentar um panorama do que seria o campo de aplicação da *Data Analytics* e de que forma ela poderia contribuir para a atuação do auditor governamental.

A fim de não encerrar o trabalho com uma análise somente do ponto de vista teórico, objetivou-se o estudo introdutório de uma ferramenta para *Data Analytics,* distinta dentre aquelas abordadas no conteúdo programático da supracitada disciplina da Especialização, e que poderia ser inserida no conjunto de habilidades de um auditor-analista de dados do setor público. Diante do exposto, o banco de dados baseado em grafos Neo4j foi escolhido para ser explorado, a fim de se obter o entendimento do valor do seu uso no contexto da auditoria governamental.

4.1 Neo4j

Apenas a análise estatística simples não consegue descrever suficientemente, muito menos prever, o comportamento dos sistemas conectados. Consequentemente, a maioria dos atuais sistemas de *Big Data Analytics* não aborda adequadamente tais

sistemas com as conexões do mundo real, ficando aquém na extração de valor e inteligência desse enorme volume de dados conectados.

Atualmente, à medida que o mundo corporativo, público ou privado, se torna cada vez mais interconectado e os sistemas cada vez mais complexos, tornou-se fundamental o uso de tecnologias construídas para destacar os relacionamentos e as características dinâmicas destes sistemas. (NEO4J, 2017)

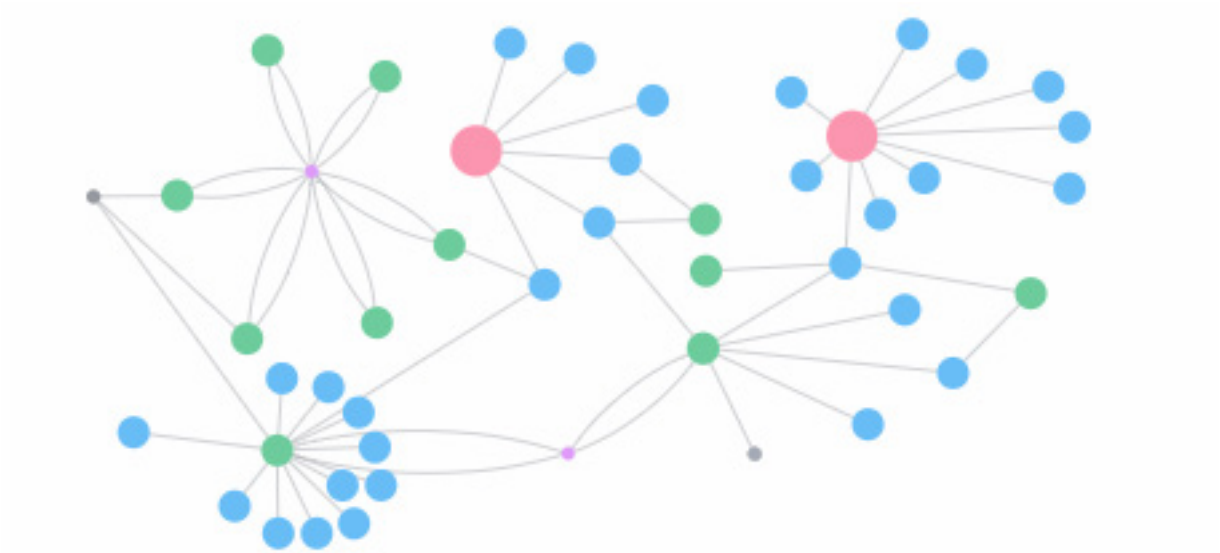
Os chamados bancos de dados relacionais - BDR foram concebidos para digitalizar dados anteriormente dispostos em formulários em papel e automatizar processos de negócios bem estruturados em formatos de tabelas. Conforme aumentam o volume de dados e a complexidade dos relacionamentos no âmbito de um BDR, verifica-se uma degradação do seu desempenho. Além disso, a adição de novos tipos de dados e relacionamentos de dados exige um novo design da arquitetura do BDR (NIXON, 2018).

Em contraponto à tecnologia dos BDR, têm-se os chamados Bancos de Dados NoSQL - *Not Only Structured Query Language* – popularmente conhecidos como Bancos de Dados não Relacionais. Estes banco de dados foram desenvolvidos para armazenar e manipular variados modelos de dados não estruturados no formato de tabelas, incluindo os seguintes modelos: documentos, chave-valor, coluna, em memória e gráficos.

No que diz respeito ao úlitmo modelo de dados supracitado, e não sendo o objetivo deste trabalho apresentar toda a complexidade por trás da intrincada teoria dos grafos, destaca-se que a finalidade de um banco de dados gráfico é facilitar a criação e a execução de aplicativos que funcionam com conjuntos de dados altamente conectados. Os casos típicos de uso de um banco de dados gráfico incluem redes sociais, mecanismos de recomendação, gráficos de conhecimento e detecção de fraudes.

Um banco de dados gráfico ou baseado em grafos amplia a capacidade de uma organização reconhecer a importância de relacionamentos e conexões persistentes entre os seus dados (ROBINSON, 2015). Neles, os relacionamentos são mais naturais, devido à própria representação gráfica, pois as entidades, chamadas de vértices (ou nós), são ligadas entre elas pelas arestas (ou relacionamentos), sendo que cada vértice ou aresta pode conter atributos próprios que os identifica, conforme a modelagem de dados estabelecida, e ainda cada relacionamento deve ter uma direção. A representação dos dados conectados no formato gráfico pode ser vista na Figura 10 (PLATFORM, 2019), em que os círculos representam os nós (vértices) e as linhas representam os relacionamentos (arestas).

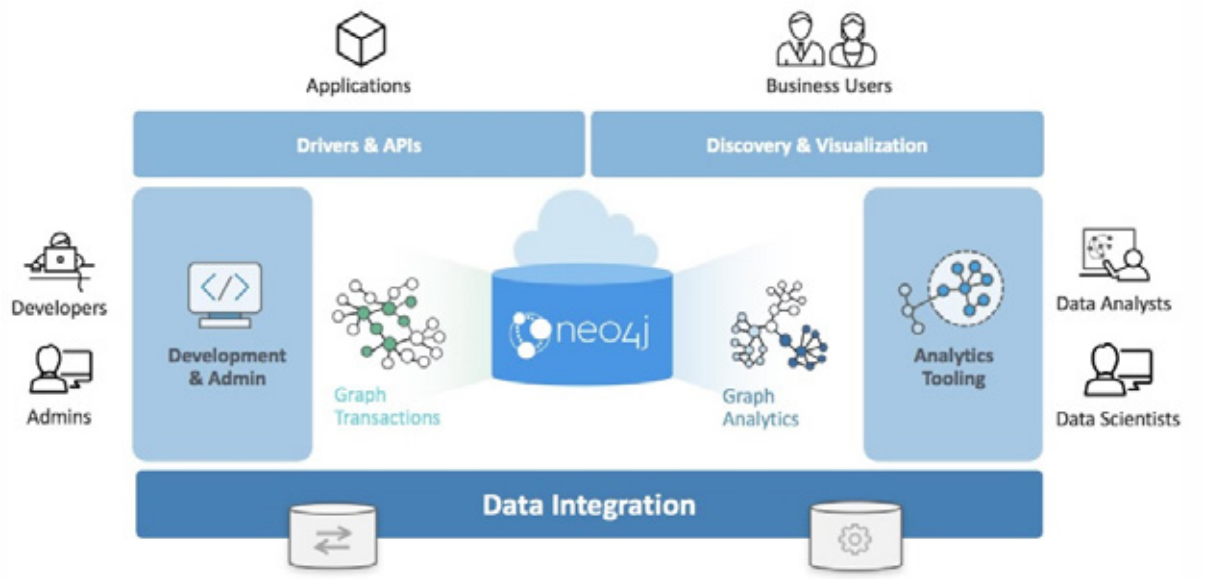
Figura 10 - Representação gráfica de dados conectados (grafos)



Entre os mais importantes bancos de dados não relacionais orientados a grafos tem-se Neo4j, desenvolvido pela empresa *Neo Tecnology*. No entanto, seu código é aberto à comunidade de pesquisa para que se possa ser modelado e aprimorado de acordo com a necessidade.

Conforme pode ser observado na Figura 11 (PLATFORM, 2019), a plataforma do Neo4j permite a integração da área de tecnologia com a de negócio de uma organização, sendo o caminho mais rápido disponível para se operacionalizar insights analíticos da atividade da organização, conectando os desenvolvedores e administradores de dados aos analistas e/ou cientistas de dados a fim de se obter novas perspecitvas bem como extrair valor desses dados.

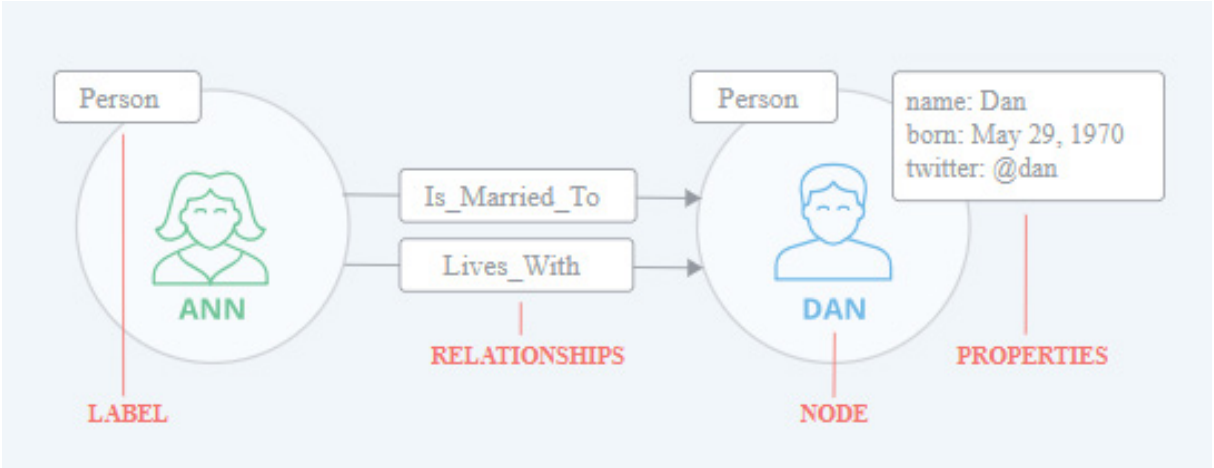
Figura 11- Plataforma gráfica do Neo4j



Esta plataforma foi desenvolvida, especificamente, para otimizar, mapear, armazenar, manipular e analisar redes de dados conectados, objetivando revelar possíveis relacionamentos ocultos, que não estariam facilmente ao alcance ao se usar um BDR.

No Neo4j os dados são armazenados nas seguintes estruturas: aresta (ou relacionamento), vértice (ou nó) e atributos (propriedades), sendo que cada nó e aresta pode ter qualquer número de atributos. Ambos nós e arestas podem ser rotulados (labels). Os rótulos podem ser usados para restringir pesquisas. Tal estruturação é apresentada na Figura 12 (PLATFORM, 2019).

Figura 12- Estrutura dos dados no Neo4j



Neste contexto, podem ser destacadas as seguintes características de cada item desta estrutura:

Nós (*Nodes*):

- São os principais elementos de dados;
- Estão conectados a outros nós por meio de relacionamentos;
- Podem ter uma ou mais propriedades (isto é, atributos armazenados como pares chave / valor)
- Têm um ou mais rótulos que descrevem seu papel no gráfico.

Relacionamentos (*Relationships*):

- Conectam dois nós;
- São direcionais;
- Os nós podem ter relacionamentos múltiplos e até recursivos;
- Podem ter uma ou mais propriedades (isto é, atributos armazenados como pares chave / valor).

Propriedades (*Properties*):

- Propriedades são valores nomeados onde o nome (ou chave) é uma string;
- Propriedades podem ser indexadas e restringidas;
- Índices compostos podem ser criados a partir de múltiplas propriedades.

Rótulos (*Labels*):

- São usados para agrupar nós em conjuntos;
- Um nó pode ter vários rótulos;
- São indexados para acelerar a localização de nós no gráfico;
- Índices de rótulos permitem otimizar a velocidade da consulta.

Não há como negar que outros sistemas de armazenamento e gerenciamento de dados têm a sua aplicabilidade em casos apropriados. No entanto, sempre que o desafio for se aproveitar das conexões entre os dados, a fim de se identificar os relacionamentos, estruturas e/ou padrões ocultos anteriormente, torna-se viável explorar a capacidade do Neo4j de encontrar novas percepções sobre como os dados de sistemas conectados operam, levando à descoberta de *insights* que podem trazer valor à organização.

Como abordado no Capítulo 2, a quantidade de dados produzidos e coletados, sobre tudo e qualquer coisa, tem crescido vertiginosamente, e as organizações muitas vezes têm encontrado problemas para armazenar, gerenciar ou até mesmo analisar tais dados. Frente à variedade dos enormes conjuntos de dados do *Big Data*, o modelo tradicional de bancos relacionais e de alguns tipos de NoSQL não conseguem descobrir as relações entre os dados e, consequentemente, não atendem às expectativas para diversas aplicações de *Data Analytics* que demandam a extração de inteligência sobre as conexões entre os dados.

Dessa forma, observa-se que um dos maiores desafios no âmbito do *Big Data* está em se relacionar os grandes volumes de dados de forma consistente e eficiente, fato este que se se torna mais desafiador quando não se sabe quais tipos de dados podem ser coletados e armazenados.

Diante do exposto, segue uma tabela comparativa, apresentando as características do Neo4j frente a outros sistemas de bancos de dados.

Características	Outros NoSQL Data Bases	Bancos Relacionais	NoSQL Graph Data Base (Ex.: Neo4j)
Armazenamento de Dados	Não há suporte para dados conectados. A confiabilidade do desempenho e dos dados diminui com a escala e a complexidade das conexões.	Armazenamento em tabelas fixas e predefinidas com linhas e colunas com dados conectados, muitas vezes separados entre tabelas, reduzindo a eficiência da consulta.	A estrutura de armazenamento resulta em transações e processamento mais rápidos para relacionamentos de dados.

Características	Outros NoSQL Data Bases	Bancos Relacionais	NoSQL Graph Data Base (Ex.: Neo4j)
Modelagem de Dados	O modelo de dados não é adequado para arquiteturas corporativas, pois colunas amplas e armazenamentos de documentos não oferecem controle no nível do design. Coloca pressão indevida no nível de aplicação para capturar e resolver problemas.	O modelo de banco de dados deve ser desenvolvido com modeladores e traduzido de um modelo lógico para um físico. Como os tipos de dados e as fontes devem ser conhecidos antes do tempo, quaisquer alterações exigem semanas de inatividade para implementação.	Modelo de dados flexível sem incompatibilidade entre o modelo lógico e físico. Os tipos e origens de dados podem ser adicionados ou alterados a qualquer momento, levando a tempos de desenvolvimento consideravelmente menores e uma iteração ágil.
Desempenho de Consulta	Nenhum recurso de processamento de gráficos para relacionamentos de dados, portanto, todos os relacionamentos devem ser criados no nível do aplicativo.	O desempenho do processamento de dados sofre com o número e a profundidade de JOINs (ou relacionamentos consultados).	O processamento de gráficos garante latência zero e desempenho em tempo real, independentemente do número ou da profundidade dos relacionamentos.
Suporte de Transação	As transações BASE levam à corrupção de dados porque a disponibilidade básica e a consistência eventual não são confiáveis para relacionamentos de dados.	Suporte a transações ACID requerido por aplicativos corporativos para dados consistentes e confiáveis.	Mantém transações ACID para dados totalmente consistentes e confiáveis o tempo todo.
Processamento em Escala	Não otimizado para ler dados em escala. A escalabilidade depende da arquitetura de scale out que não protege a integridade de dados semelhantes a gráficos, portanto, os dados não são confiáveis.	Escalar através da replicação e dimensionar a arquitetura é possível, mas caro. Os relacionamentos de dados complexos não são coletados em escala.	Modelo dimensionado para consultas baseadas em padrões. A arquitetura de scale out mantém a integridade dos dados por meio da replicação. Possibilidades de escalonamento massivo.
Linguagem de Consulta	A linguagem de consulta varia, mas não existem construções de consulta para expressar relacionamentos de dados.	SQL: Uma linguagem de consulta que aumenta em complexidade com o número de JOINs necessários para consultas de dados conectadas.	Cypher: Uma linguagem de manipulação e consulta baseada em grafo, que fornece uma forma eficiente e expressiva de descrever os relacionamentos.

Comparativo entre Neo4j e outros Sistemas de Bancos de Dados (PLATFORM, 2019)

4.2 *Cypher*

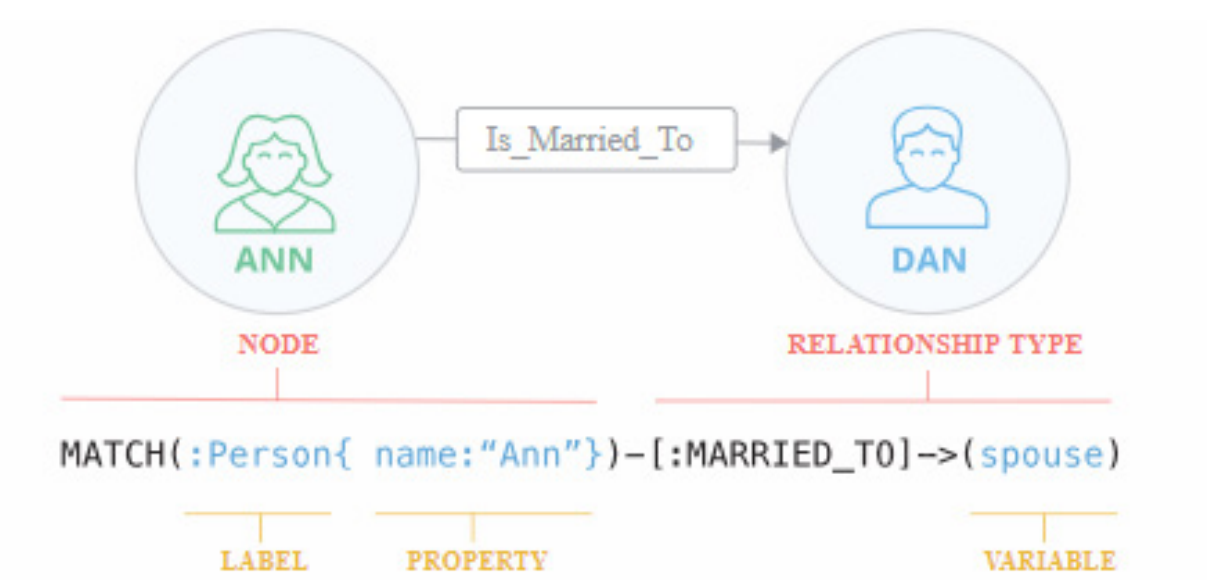
A linguagem de manipulação e consulta padrão para o banco de dados baseado em grafos Neo4j se chama *Cypher*. Ela foi originalmente destinada a ser usada com o Neo4j, no entanto, o seu código foi aberto para que a comunidade de pesquisadores e estudantes pudesse contribuir no seu desenvolvimento e também visando a sua popularização.

Com a *Cypher* são criados os elementos padrões de dados gráficos: nós e relacionamentos. Normalmente, os nós são criados com um ou mais rótulos, que são informações sobre estes nós para fins de agregação nas consultas, e podem ter relacionamentos múltiplos e/ou até recursivos, conforme apresentou a Figura 12. Ainda, nós e relacionamentos também têm zero ou mais propriedades, na qual uma propriedade é uma ligação de valor-chave no formato de *string.*

A *Cypher* é uma linguagem de consulta declarativa com sintaxe inspirada na SQL e permite aos seus usuários declarar quais ações eles querem executar (como consultar, inserir, atualizar ou excluir) sobre o seu modelo de dados baseado em grafos. Neste contexto de similaridade com a SQL, que se tornou padrão no âmbito dos bancos de dados relacionais, uma vez que a *Cypher* é altamente legível, após o seu aprendizado ela também é fácil de manter, simplificando a manutenção da aplicação (CYPHER, 2019).

A Figura 13 (CYPHER, 2019) apresenta a sintaxe padrão de uma instrução básica em *Cypher,* que busca um padrão de relacionamento entre dois nós, destacando os principais elementos de um banco de dados baseado em grafos como o Neo4j.

Figura 13- Instrução em Cypher no Neo4j



Analisando a supracitada instrução, as principais informações a se lembrar, quando da escrita das instruções, são:

- Os nós são representados por parênteses ( ), lembrando os círculos da representação gráfica;
- A direção de um relacionamento é representado por “setas” ->;
- Informações sobre um relacionamento devem ser inseridas entre colchetes [ ];
- As propriedades são inseridas entre { }.

As consultas mais simples em *Cypher* consistem de uma cláusula *MATCH* seguida por uma cláusula *RETURN*. A cláusula *MATCH* permite especificar os padrões que o Neo4j irá procurar no banco de dados. Essa é a principal maneira de obter as conexões dos conjuntos de dados no conjunto. Já a *RETURN* especifica quais nós, relacionamentos e/ou propriedades nos dados correspondentes devem ser devolvidos ao usuário da consulta.

A fim de demonstrar a similaridade da *Cypher* com a SQL, seguem algumas cláusulas importantes na linguagem de banco de dados baseados em grafos:

- *WHERE*: Fornece critérios para filtrar resultados;
- *CREATE*: Cria nós e relacionamentos;
- *MERGE*: Garante que um padrão exista no gráfico ou que precisa ser criado;
- *DELETE*: Remove nós, relacionamentos e propriedades;
- *SET:* Define os valores da propriedade;
- *FOREACH:* Executa uma atualização para cada elemento em uma lista;
- *UNION:* Mescla resultados de duas ou mais consultas.

É fato que o ciclo de aprendizado da linguagem não é tão básico, mas existe uma ampla documentação disponibilizada pela empresa desenvolvedora do Neo4j. Além disso, como o projeto da *Cypher* foi aberto, a linguagem está em uma rápida evolução e a cada nova versão publicada diversas melhorias e funcionalidades são adicionadas para que os usuários possam desenvolver suas aplicações e análises na busca de encontrar relacionamentos entre os dados.

Assim, *Cypher* e Neo4j apresentam-se que como um conjunto de ferramentas que podem ser utilizadas pelos auditores governamentais, não necessariamente especialistas em tecnologia da informação, nos seus trabalhos de descoberta de fraude, por exemplo.

Embora nenhuma medida de prevenção de fraude possa ser perfeita, uma oportunidade significativa de melhoria pode ser obtida olhando além dos pontos de dados individuais, ou seja, voltar as atenções para as conexões que os ligam. Muitas vezes essas conexões passam despercebidas até que seja tarde demais, o que é lamentável, pois tais conexões muitas vezes contêm os melhores indícios ou mesmo as próprias evidências. Entender as conexões entre os dados e obter o significado desses links pode trazer insights significativos no âmbito de uma auditoria, que seriam difíceis de se detectar caso a análise fosse empreendida em representações de dados em forma de tabelas (RATHLE, 2017).

Portanto, implementar projetos de análise em bancos de dados baseados em grafos no contexto das auditorias governamentais, nesta era do *Big Data*, pode contribuir para se resolver uma variedade de problemas em bases de dados conectados, incluindo a detecção de fraudes decorrentes do relacionamento entre agentes públicos e privados.

4.3 Execução do experimento

O objetivo do presente tópico é avaliar a experiência da implementação de um modelo de banco de dados orientado a grafos, por meio do Neo4j, a fim de vislumbrar a possibilidade de aplicação dessa metodologia em trabalhos de auditoria governamental.

Incialmente, destaca-se que a característica que define um banco de dados orientado a grafos é o uso de um modelo que representa suas estruturas de dados por meio de nós e arestas (ou relacionamentos, no contexto do Neo4j).

No modelo lógico de um banco de dados orientado a grafos há a representação visual explícita do relacionamento entre os dados, que por sua vez, no modelo relacional, somente se torna visível por meio da elaboração de JOINs em SQL que exigem, conforme o volume de dados, bastante capacidade de processamento.

Já a tecnologia por trás do Neo4j implementa a chamada arquitetura nativa de grafos, usando listas de adjacência livres de índices, teoria que se refere à capacidade de se encontrar todos os nós adjacentes a um nó qualquer, sem requerer uma busca em toda a base de dados. Um nó possui uma referência aos nós adjacentes por meio das arestas e, por sua vez, cada aresta está referenciada ao nó inicial e ao nó final. Dessa forma, o chamado custo do caminho entre todos nós adjacentes é diretamente proporcional ao número de relacionamentos aos quais está vinculado.

Adicionalmente, dependendo da questão a ser respondida pelo tipo de consulta ao banco de dados orientado a grafos, pode ser necessário filtrar os relacionamentos pelos seus atributos ou sua direção. No contexto em que os nós são bastante relacionados, ou conectados, a performance esperada para cada consulta é muito pior, demandando grande capacidade computacional.

Nesse sentido, como o presente trabalho buscou explorar uma nova tecnologia para ser agregada ao ferramental de trabalho de um auditor governamental, a metodologia do experimento visou simplificar o modelo de dados implementado, de forma que possíveis limitações de infraestrutura tecnológica pudessem ser minimizadas.

Diante do exposto, a fonte dos dados para a criação do modelo foi extraída no site do Portal da Transparência do Governo Federal, disponível no seguinte endereço *http://transparencia.gov.br/download-de-dados/compras*. Tal endereço fornece variados dados de compras governamentais por ano e mês, sendo possível extrair informações no formato de tabela de três arquivos principais com extensão .csv, quais sejam: “Compras”, “ItemCompra” e “TermoAditivo”.

Em uma exploração inicial básica nos dicionários de dados de cada arquivo, vislumbrou-se variadas possibilidades para implementação do modelo, com a possibilidade de junções entre as colunas dos arquivos. No entanto, tendo como foco a minimização de possíveis limitações tecnológicas ainda desconhecidas no início da modelagem, e levando também em consideração o estágio inicial do conhecimento e aprendizado do Neo4j, optou-se por usar somente a fonte de dados “Compras.csv”, que apresentava as seguintes colunas de dados:

“Número do Contrato”; “Objeto”; “Fundamento Legal”; “Modalidade Compra”; “Situação Contrato”; “Código Órgão Superior”; “Nome Órgão Superior”; “Código Órgão”; “Nome Órgão”; “Código UG”; “Nome UG”; “Data Assinatura Contrato”; “Data Publicação DOU”; “Data Início Vigência”; “Data Fim Vigência”; “CNPJ Contratado”; “Nome Contratado”; “Valor Inicial Compra”; “Valor Final Compra”.

Inicialmente, a expectativa era criar um modelo de dados abrangendo todo um exercício financeiro. No entanto, ao se efetuar o download de todos os meses do exercício de 2017 e ao carregar parcialmente tais dados no Banco de Dados Relacional MySQL, verificou-se um volume muito grande de compras no exercício, o que resultaria em milhares de nós e, possivelmente, milhões de relacionamentos para serem modelados.

Então, a fim de atingir o objetivo de explorar uma ferramenta de *Data Analytics* passível de uso pelo auditor governamental, e visando evitar o risco de não implementação do modelo proposto, dentre outros fatores, devido à provável degradação do desempenho do equipamento de suporte computacional usado no experimento, a

fonte de dados para análise, tratamento e modelagem ficou restrita aos registros do mês de janeiro, do exercício de 2017.

Em um processo de exploração dos dados do mês de janeiro de 2017 no MySQL, verificou-se que no período foram realizadas 3.754 compras que totalizaram o montante de R$ 4.542.965.287,29, sendo este o resultado da seguinte consulta SQL:

Consulta 1:

SELECT nome_orgao, data_publicacao_dou, nome_contratado, valor_final_compra FROM BD.201701_Compras ORDER BY 1, 3;

Os termos respectivos nome_orgao, data_publicacao_dou, nome_contratado, valor_final_compra, referem-se às colunas “Nome Órgão”, “Data Publicação DOU”; “Nome Contratado” e “Valor Final Compra” do arquivo “Compras.csv”.

O próximo passo foi identificar a quantidade de órgãos e/ou entidades e de empresas que transacionaram esse número de compras, já que esses seriam os nós do modelo baseado em grafos. Para tanto foram realizadas as seguintes consultas SQL:

Consulta 2:

SELECT nome_orgao, count(nome_orgao) as qtd_ocorrencia

FROM BD. 201701_Compras GROUP BY nome_orgao ORDER BY 2 DESC;

Consulta 3:

SELECT nome_contratado, count(nome_contratado) as qtd_ocorrencia

FROM BD. 201701_Compras GROUP BY nome_contratado ORDER BY 2 DESC;

A consulta 2 retornou todos os órgãos da base de dados que realizaram pelo menos uma compra e a quantidade de ocorrências deles na base, ou seja, a quantidade de compras que cada um fez. O resultado da consulta é que 211 órgãos/entidades fizeram compras no período, sendo que 27 órgãos só fizeram uma compra e 184 fizeram mais de uma compra. Dentre o conjunto de órgãos, os 10 órgãos que mais compraram fizeram 1.954 compras, aproximadamente 52,05% do total de compras, perfazendo um montante contratado de R$ 1.790.290.946,12, o que representou 39,41% das contratações do período.

Já a consulta 3 trouxe as empresas contratadas e a quantidade em que cada uma foi contratada. No universo de 3.754 compras, 2.286 empresas venderam produtos ou

serviços para os 211 órgãos/entidades. Do total de empresas, 525 fizeram mais de uma venda e 1.761 fizeram somente uma venda, em janeiro de 2017.

Em seguida às análises das três supracitadas consultas, foi realizada a exportação, para o Microsoft Excel, dos dados da primeira consulta, para fins manipulação e análise dos dados. No Excel, em uma planilha auxiliar, foi gerada uma coluna com um par ordenado (nome_orgao, nome_contratado) e outra coluna com o respectivo "valor_final_compra" de cada par ordenado, representando cada uma das 3.754 compras efetuadas pelos 211 órgãos.

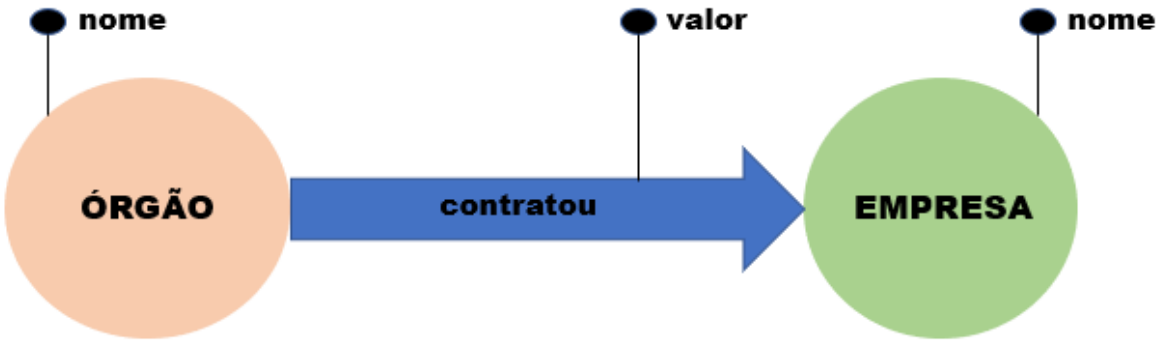
Em seguida, fez-se o uso da função SUBTOTAL para realizar a soma total da coluna "valor_final_compra" de cada par ordenado. Por fim, usou-se a função FILTRO para restringir a exibição de dados somente aos totais calculados.

O resultado de todas etapas executadas no Excel foi uma planilha contendo 2.745 pares ordenados (nome_orgao, nome_contratado), sem qualquer ocorrência duplicada, e o respectivo valor de cada par ordenado, valor este que representa o somatório total calculado para as compras efetuadas, no mês de referência, entre um órgão e uma determinada empresa. Estes 2.745 registros únicos, nada mais são que o total de relacionamentos observados entre órgãos e empresas no período.

Diante do exposto, o modelo com 2.745 relacionamentos a serem implementados, a partir dos dados utilizados, seria composto por 211 nós, representando os órgãos/entidades contratantes, e 2.286 nós, representando as empresas contratadas, totalizando 2.497 nós.

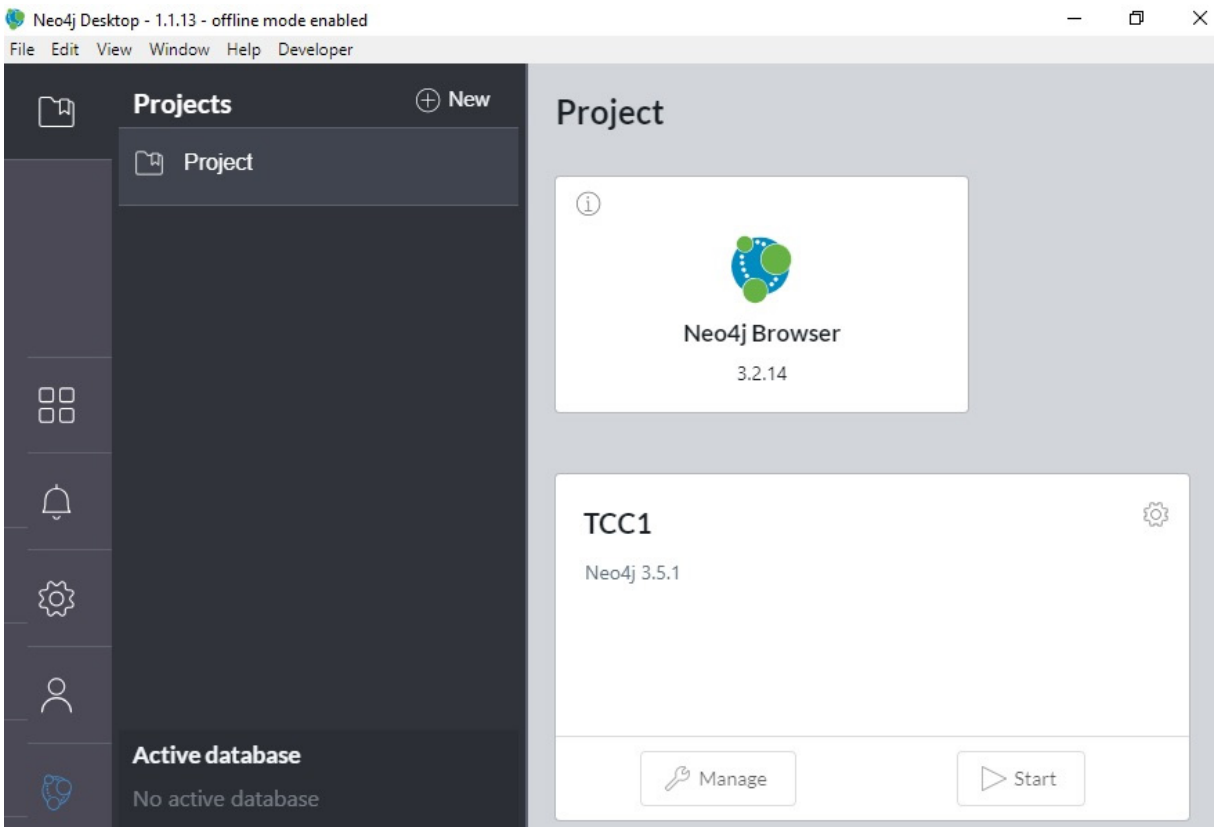
Para fins de simplificação da implementação do modelo, à cada nó e relacionamento foi atribuída somente uma propriedade, conforme figura 14.

Figura 14- Modelo de conexão entre os nós.



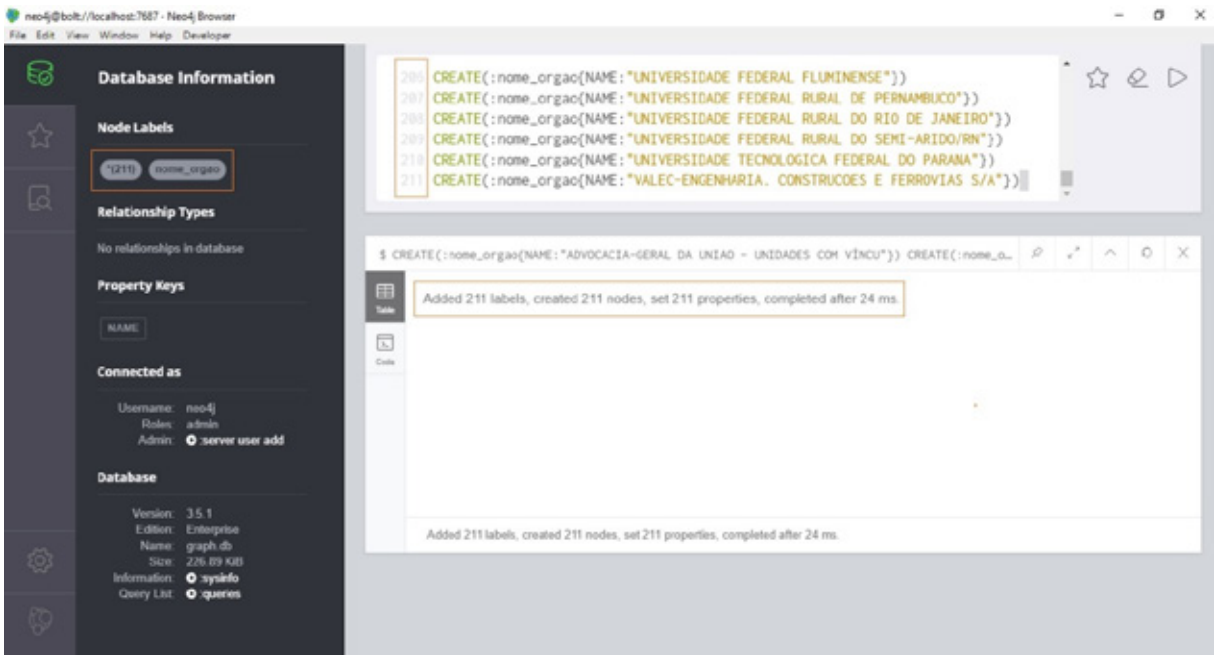
A partir das diretrizes estabelecidas, foi iniciado o processo de construção do projeto denominado TCC1 no Neo4J, conforme apresentado a seguir:

Figura 15- Criação do projeto para implementação do modelo no Neo4j



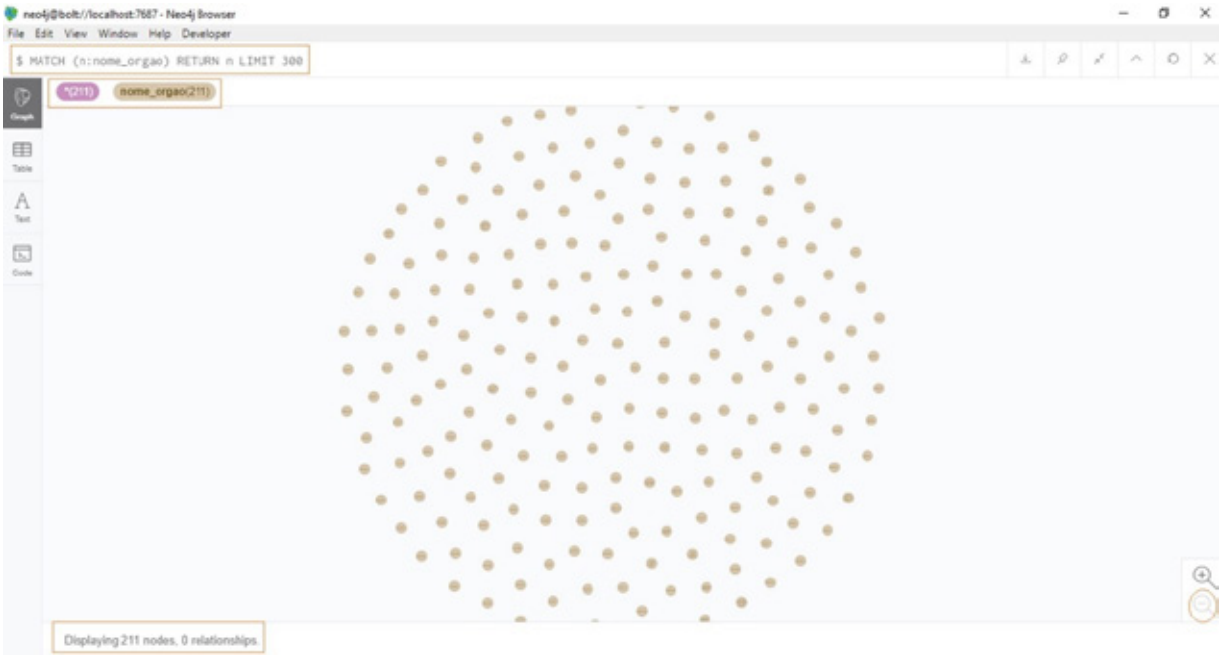
Inicialmente, foram criados os nós referentes aos 211 órgãos. A instrução em Cypher para criar cada nó foi: **CREATE**(:nome_orgao{NAME:"NODE_ORGAO"}).

Figura 16- Criação dos nós dos órgãos/entidades.



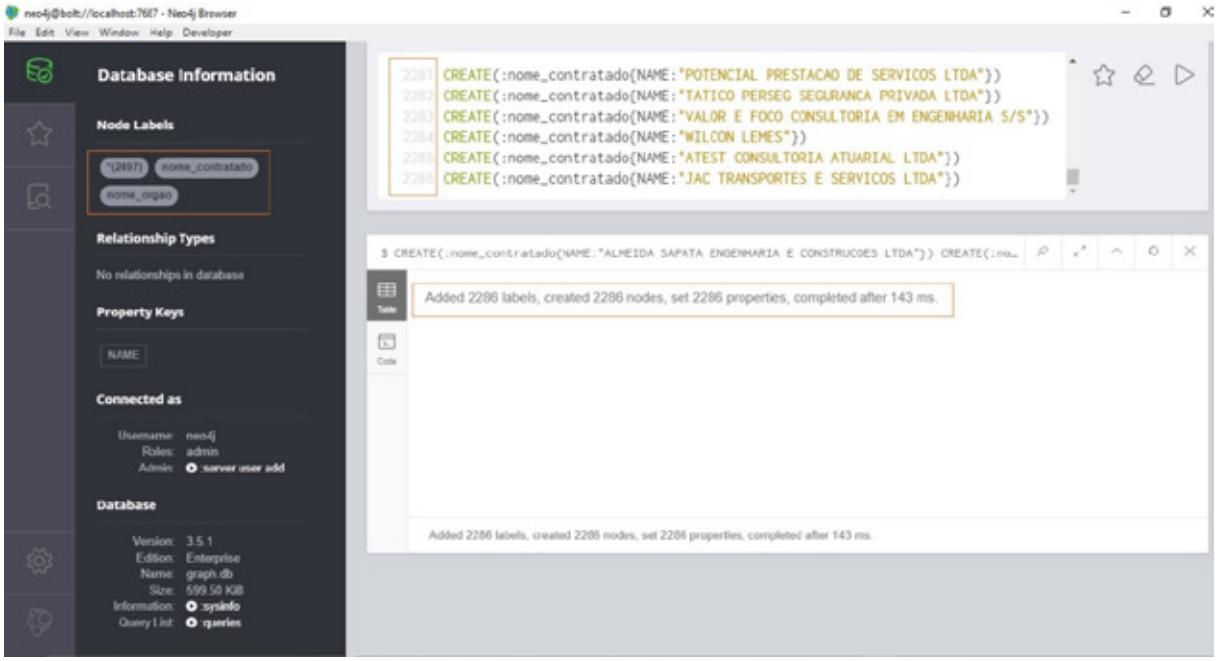
A execução de 211 instruções, conforme sintaxe apresentada, resultou no seguinte padrão gráfico, representando todos os órgãos:

Figura 17- Conjunto de nós criados para os órgãos/entidades



Em seguida, foram criados os nós das 2.286 empresas contratadas, com a seguinte sintaxe por nó: **CREATE**(:nome_contratado{NAME:"NODE_CONTRATADO"}).

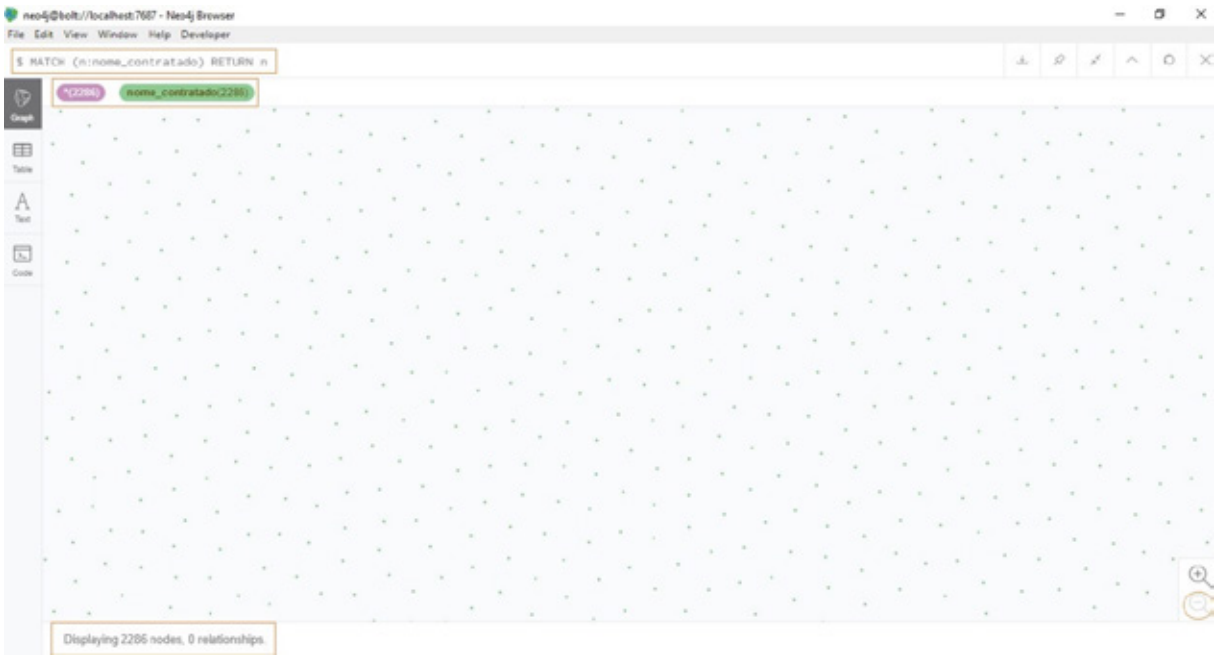
Figura 18- Criação dos nós das empresas contratadas



Em relação às figuras 16 e 18, percebe-se a implementação do modelo com dois tipos de Rótulos (*Labels*): "nome_orgao" e "nome_contratado". No Neo4j os *Labels* permitem agrupar nós em conjuntos, o que favorece a localização de nós no gráfico e, consequentemente, incrementa a velocidade da consulta a ser realizada.

Como a quantidade de nós para as empresas era superior em quase 10 vezes a quantidade de nós para órgãos, uma representação com qualidade, do padrão gráfico de todos esses nós, não foi possível, conforme pode ser observado na figura 19 a seguir.

Figura 19- Conjunto de nós criados para as empresas



Por conseguinte, a última fase para a implementação do modelo apresentado na figura 14 foi a criação dos 2.745 relacionamentos. As instruções em *Cypher* inseridas no Neo4j apresentam a seguinte sintaxe:

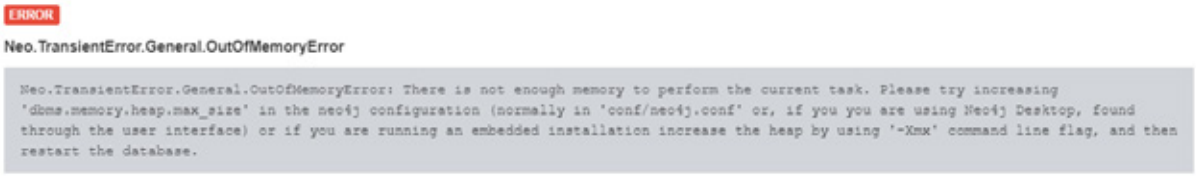
MATCH (a:nome_orgao), (b:nome_contratado)

WHERE a.NAME = “NODE_ORGAO” AND b.NAME = “NODE_CONTRATADO”

CREATE (a)-[:CONTRATOU{valor:”VALOR_DO_RELACIONAMENTO”}]->(b)

Destaca-se que no momento de implementação dos relacionamentos foram encontrados problemas na capacidade de processamento do equipamento usado para desenvolver o experimento prático. Não foi possível carregar, concomitantemente, as 2.745 instruções para criação dos relacionamentos. Neste tipo de tentativa, o equipamento usado apresentou problemas de memória e, mesmo com a alternativa apresentada pelo Neo4J, conforme figura 20, não foi possível sanar o problema.

Figura 20- Problema no carregamento dos relacionamentos



Segundo informações da comunidade de desenvolvimento do Neo4j, o estado de toda transação deste banco de dados é mantido na memória do computador/servidor, portanto, se a inserção do conjunto de dados estiver em uma única transação/instrução, isso pode causar problemas, como o estouro de memória verificado. Portanto, optou-se pela inserção das instruções de criação dos relacionamentos em grupos de 200 instruções, aproximadamente, fato este que permitiu a criação de todos relacionamentos do modelo. Dessa forma, por questões de praticidade frente às limitações encontradas, mostrou-se adequada a restrição do uso de mais propriedades para os nós e os relacionamentos. A figura 21 apresenta a criação dos relacionamentos necessários.

Figura 21- Criação dos relacionamentos entre órgãos/entidades e empresas.

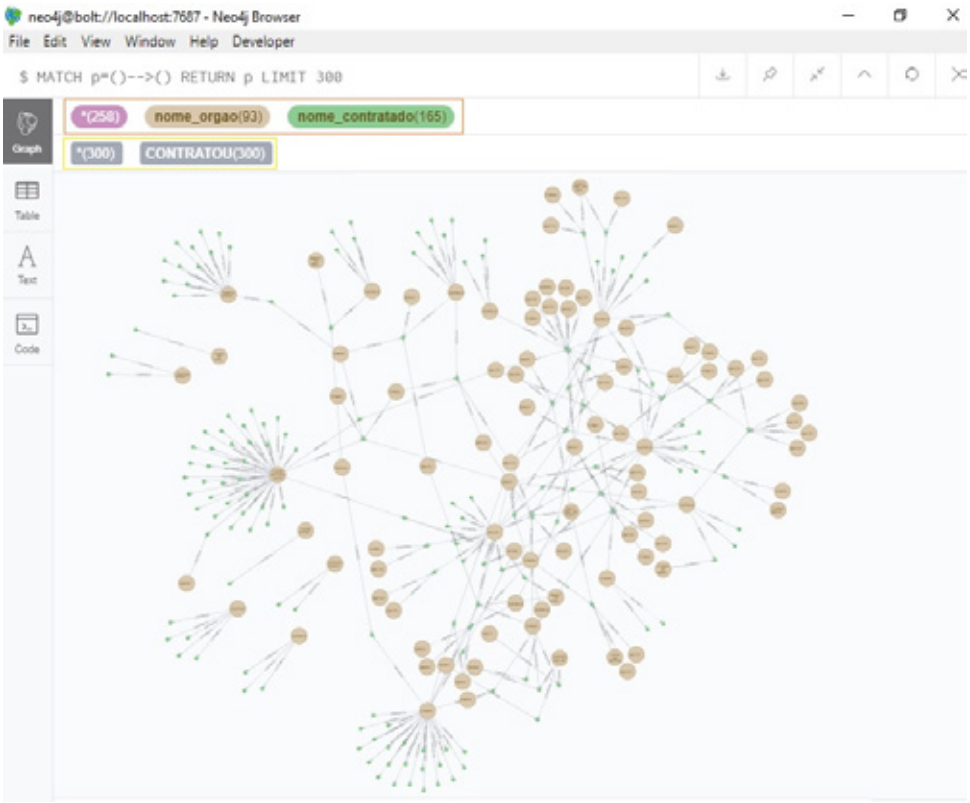
Cada aresta, que é a representação gráfica do relacionamento, foi criada com o tipo "CONTRATOU". Quando da visualização gráfica dos nós e relacionamentos, ao se clicar sobre o relacionamento/aresta é possível visualizar os seus atributos que, neste caso prático, foi somente a propriedade "valor".

A figura 22 apresenta, parcialmente, haja vista a quantidade de nós criados, todos os relacionamentos entre os nós órgãos e empresas.

Figura 22-Representação parcial dos relacionamentos entre todos os nós

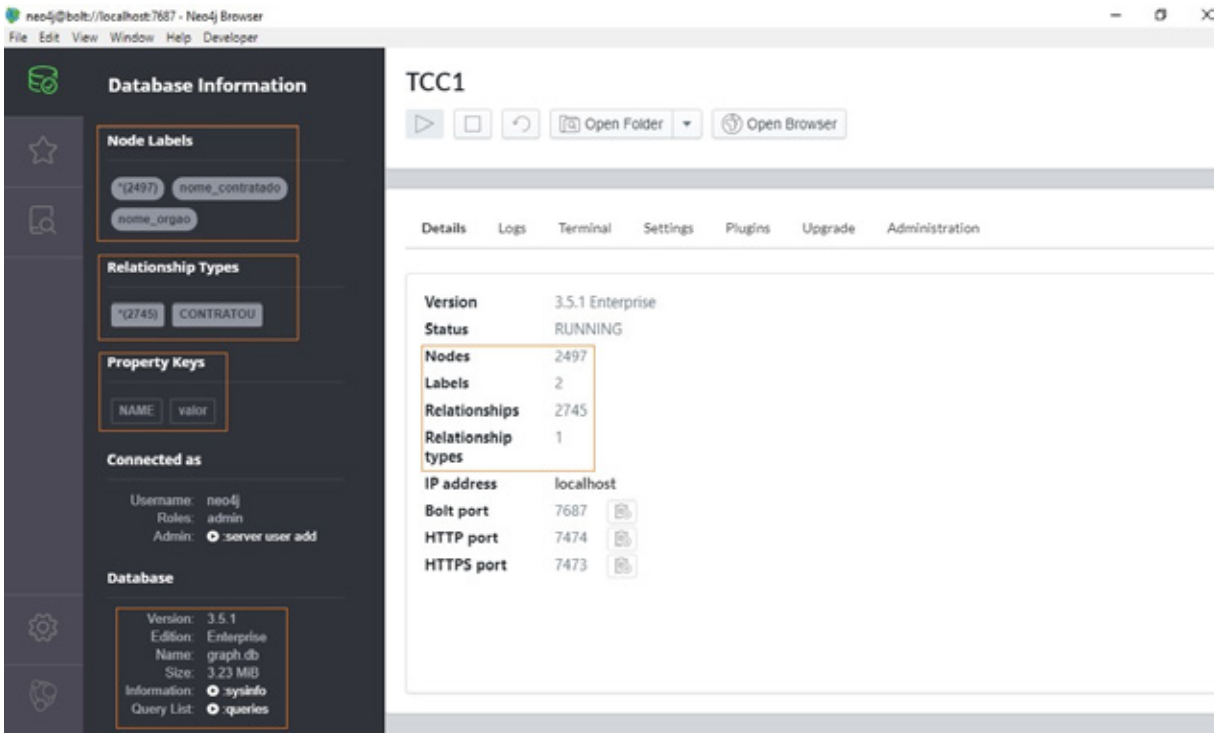
Conforme representa a figura 23, é possível ao analista dos dados limitar a quantidade de relacionamentos, a serem visualizados, conforme a necessidade do agrupamento a ser feito para fins de análise.

Figura 23-Limitando a quantidade de relacionamentos visualizados



Após a implementação do modelo, é possível identificar no Neo4j todas as principais informações de um banco de dados baseados em grafos: nós, relacionamentos, rótulos e propriedades, conforme figura 24.

Figura 24- Informações do Modelo de Banco de Dados Gráfico



4.4 Interpretação do experimento

Pensando em uma situação de aplicação de procedimento de auditoria, quando o objetivo ou necessidade for a análise individual de um órgão jurisdicionado, a modelagem gráfica do Neo4j pode se apresentar como uma ferramenta para que o auditor possa obter *insights* da base de dados sob escrutínio, extraindo valor das relações deste órgão com outras pessoas, físicas ou jurídicas.

A figura 25 foi apresentada para representar a seleção de um órgão específico e seus relacionamentos. A partir da consulta é possível fazer a análise individual do órgão, verificando com quem este mais transacionou no período e qual foi montante de tal conexão.

Evidentemente, como o modelo implementado no presente experimento teve a complexidade das propriedades dos nós e relacionamentos limitada, em uma situação real mais variáveis devem estar sob o foco da análise do auditor para que haja qualidade e efetividade aos *insights* descobertos.

Figura 25 - Análise focada em um órgão e seus relacionamentos (adaptada)

Nesse sentido, como a fonte de dados do estudo contém registros das compras governamentais em determinado período, em um caso real com maior complexidade técnica, o auditor poderia inserir no modelo baseado em grafos mais atributos aos relacionamentos, por exemplo: “Objeto”; “Modalidade Compra”; “Data Início Vigência”; “Data Fim Vigência”; “Valor Inicial Compra”; “Valor Final Compra”.

Com mais informações para análise, o auditor governamental pode buscar melhores *insights* a fim de compreender os motivos pelos quais um órgão compra demasiadamente determinado produto ou as razões pelas quais uma entidade contrata periodicamente serviços sob a ótica da inexigibilidade, dentre outras possibilidades, tudo do ponto de vista gráfico.

Outra possibilidade de aplicação do Neo4j, em um caso prático, seria para analisar aqueles citados 10 órgãos que mais compraram em determinado período. Como a materialidade das contratações representou, aproximadamente, 40% do total contratado, e como os recursos, materiais e humanos, disponíveis aos órgãos de controle não permitem avaliar/fiscalizar todo o universo das transações públicas, uma boa solução seria focar nas conexões dos maiores contratantes.

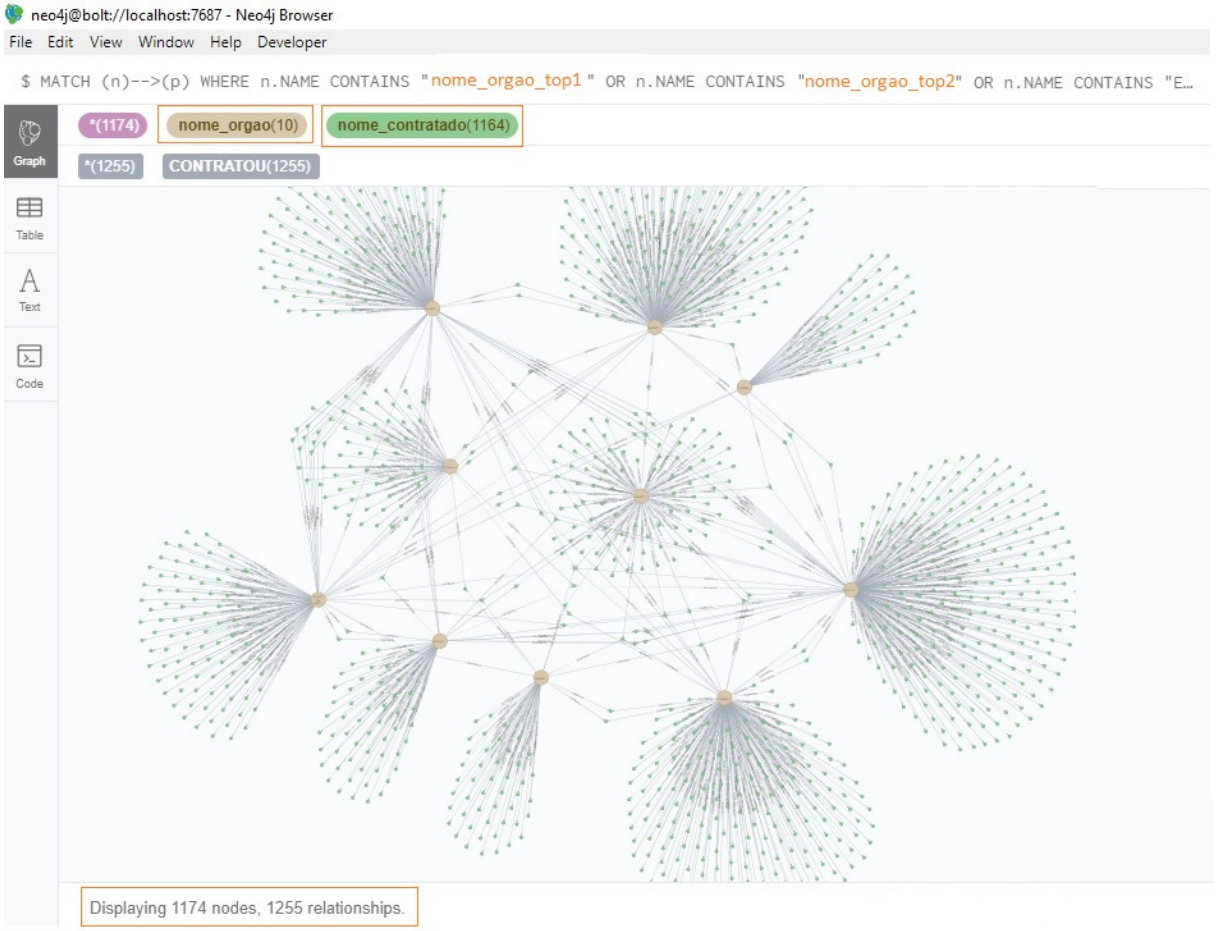
É fato que não há garantia nessa abordagem, no entanto, faz sentido direcionar recursos de auditoria escassos para aquelas situações que apresentam maior quantidade de relacionamentos entre órgãos, entidades, empresas e/ou pessoas e com maior montante financeiro envolvido. A figura 26, precedida pela instrução *Cypher* que gerou

as conexões dos 10 órgãos que mais contrataram no período em análise, ilustra esta última situação prática destacada. Destaca-se que tais órgãos foram identificados no momento do pré-processamento dos dados no *MySQL.*

Instrução *Cypher:*

MATCH (n)-->(p) WHERE n.NAME CONTAINS “nome_orgao_TOP1” OR n.NAME CONTAINS “ nome_orgao_TOP2” OR n.NAME CONTAINS “ nome_orgao_TOP3” OR n.NAME CONTAINS “ nome_orgao_TOP4” OR n.NAME CONTAINS “ nome_orgao_TOP5” OR n.NAME CONTAINS “ nome_orgao_TOP6” OR n.NAME CONTAINS “ nome_orgao_TOP7” OR n.NAME CONTAINS “ nome_orgao_TOP8” OR n.NAME CONTAINS “ nome_orgao_TOP9” OR n.NAME CONTAINS “ nome_orgao_TOP10” RETURN n, p

Figura 26 - Dez órgãos que contrataram, aproximadamente, 40% do montante do período.



5. Conclusão

A aplicação de ferramentas de *Data Analytics* no conjunto dos dados coletados e armazenados pelos órgãos e entidades governamentais tem um grande potencial para impactar, positivamente, a realidade de atuação dos órgãos de controle.

No entanto, para que haja o desenvolvimento contínuo do campo da Análise de Dados como subsídio ao controle e fiscalização da atuação governamental, é fundamental que a metodologia de *Data Analytics* passe a fazer parte da estratégia de negócio dos órgãos encarregados do controle dos recursos públicos.

Dessa forma, o apoio e incentivo da alta gestão dos órgãos de controle, por meio de ações de investimento em melhoria de processos, em infraestrutura tecnológica e em capacitação, é essencial para que o processo análise dos dados da administração pública possa gerar valor, não somente no âmbito de atuação da auditoria governamental, mas também para a sociedade que confia os seus recursos ao gestor público.

Este trabalho se propôs a apresentar uma abordagem sobre o campo de aplicação da *Data Analytics*, destacando o banco de dados orientado a grafos Neo4j, apresentando-o como possível ferramenta a ser aplicada à atuação do auditor governamental.

A partir da revisão teórica da literatura e da implementação da modelagem básica de uma situação prática da administração pública, por meio do experimento foi possível concluir que o Neo4j, e seu modelo orientado a grafos, se apresenta como uma ferramenta de *Data Analytics* adequada ao emprego dos órgãos de controle.

No âmbito de uma auditoria governamental, a possibilidade de aplicação de um banco de dados orientado a grafos se destaca, principalmente, naquelas ocasiões nas quais o relacionamento entre os dados é tão ou mais importante do que os próprios dados para a tomada de decisões.

Por fim, considerando a motivação do presente trabalho de conclusão de curso da Especialização, sugiro, como trabalho futuro, um experimento com o uso do Neo4j em uma Auditoria Financeira para modelar as conexões entre o ente auditado e as partes relacionadas destacadas nas notas explicativas das demonstrações financeiras, sejam elas pessoas físicas ou jurídicas, ou para encontrar os padrões de relacionamentos entre os dirigentes do ente e os dirigentes das outras pessoas jurídicas com as quais ele assumiu obrigações financeiras. Nessas ocasiões, a análise dos dados por meio de uma modelagem no Neo4j pode contribuir para detectar situações de conflito de interesse ou mesmo de fraude e/ou corrupção.

Referências Bibliográficas

BONFIM, Marcus V. de Jesus; Transparência e accountability na comunicação pública: impactos da Lei de Acesso à Informação nos órgãos públicos paulistas. Tese (Mestrado em Ciências da Comunicação) - Escola de Comunicação e Artes da USP, P. 13. 2015.

CGU. Ferramenta para avaliação preventiva e automatizada de editais de licitação, 2015. Disponível em: <https://www.cgu.gov.br/noticias/2015/06/controladoria-lanca-ferramenta-para-avaliacao-preventiva-e-automatizada-de-editais-de-licitacao>. Acesso em: 25 de abril de 2019.

CHILE. Norma Técnica para Publicación de Datos Abiertos en Chile, 2013. Disponível em: <http://instituciones.gobiernoabierto.cl/NormaTecnicaPublicacionDatosChile_v2-1.pdf>. Acesso em: 16 de fevereiro de 2019.

CHUI, Michael et al. Big data: The Next Frontier For Innovation, Competition, And Productivity, 2011. Disponível em: < https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation>. Acesso em: 16 de fevereiro de 2019.

COSTA, Gledson P. Corrêa; Contribuições da Auditoria Contínua para a efetividade do Controle Externo. Trabalho de Conclusão de Curso (Especialização em Auditoria e Controle Governamental) – Instituto Serzedello Corrêa – ISC/TCU. 2012.

CYPHER, Language. Cypher: The Standard Query Language for Graph Database Technology, 2019. Disponível em: < https://neo4j.com/cypher-graph-query-language/>. Acesso em: 6 de março de 2019.

DELOITTE. Analytics na auditoria interna - Direcionamento inteligente rumo a 2020, 2016. Disponível em: <https://www2.deloitte.com/content/dam/Deloitte/br/Documents/risk/POV_analytics_auditoria_interna.pdf>. Acesso em: 16 de fevereiro de 2019.

FROST, Frost & Sulivan. Latin American Big Data and Analytics (BDA) market, 2018. Disponível em: <https://ww2.frost.com/news/press-releases/brasil-e-mexico-se-destacam-na-adocao-de-big-data-na-america-latina-afirma-frost-sullivan/>. Acesso em: 16 de fevereiro de 2019.

GANDOMI, Amir; HAIDER, Murtaza. Beyond the hype: Big data concepts, methods, and analytics, 2015.Disponível em: <https://www.sciencedirect.com/science/article/pii/S0268401214001066 >. Acesso em: 16 de fevereiro de 2019.

GANTZ, John; REINSEL, David. The Digital Universe In 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. EMC Corporation, 2012. Disponível em: <https://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf >. Acesso em: 16 de fevereiro de 2019.

HADOOP. Apache Hadoop, 2019. Disponível em: < https://hadoop.apache.org/>. Acesso em: 6 de março de 2019.

HAMMER, Cornelia; KOSTROCH, Diane; QUIROS, Gabriel. Big Data: Potential, Challenges, and Statistical Implications, 2017. Disponível em: <https://www.imf.org/en/Publications/Staff-Discussion-Notes/Issues/2017/09/13/Big-Data-Potential-Challenges-and-Statistical-Implications-45106>. Acesso em: 6 de março de 2019.

IBM. Mitos sobre o Big Data, 2013. Disponível em: <https://www.ibm.com/developerworks/community/blogs/ctaurion/entry/mitos_sobre_big_data?lang=en>. Acesso em: 6 de março de 2019.

IBM. Apache Hadoop: Built for big data, insights, and innovation, 2019. Disponível em: <https://www.ibm.com/analytics/hadoop>. Acesso em: 6 de março de 2019.

IDC (International Data Corporation). Worldwide Semiannual Big Data and Analytics Spending Guide, 2018. Disponível em: <https://www.idc.com/url.do?url=/includes/pdf_download.jsp?containerId=prUS44215218&position=1>. Acesso em: 16 de fevereiro de 2019.

KAPPAL, Sunil. R vs. Python, 2017. Disponível em: <https://dzone.com/articles/r-or-python-data-scientists-delight>. Acesso em: 6 de março de 2019.

KAUFMAN, Marcia et al. Big Data for Dummies. New Jersey, 2013.

KING, John; MAGOULAS, Roger. 2016 Data Science Salary Survey, 2016. Disponível em: < https://www.oreilly.com/data/free/files/2016-data-science-salary-survey.pdf>. Acesso em: 6 de março de 2019.

KUENKAIKAEW, Siripan. Predictive Audit Analytics: Evolving to a new era, 2013. Disponível em: < https://rucore.libraries.rutgers.edu/rutgers-lib/41494/PDF/1/play/>. Acesso em: 16 de fevereiro de 2019.

LANEY, Doug. 3D Data Management: Controlling Data Volume, Velocity and Variety, 2001. Disponível em: <https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>. Acesso em: 16 de fevereiro de 2019.

LETOUZÉ, Emmanuel; JÜTTING, Johannes. Official Statistics, Big Data and Human Development: Towards a New Conceptual and Operational Approach, 2014. Disponível em: <https://www.odi.org/sites/odi.org.uk/files/odi-assets/events-documents/5161.pdf>. Acesso em: 6 de março de 2019.

MARQUES, Pedro M. B. Técnicas de Análise de Dados (Data Analytics) no contexto de uma auditoria financeira (PARTE 1), 2016. Disponível em: <http://www.oroc.pt/fotos/editor2/Revista/73/Auditoria.pdf>. Acesso em: 6 de março de 2019.

MARR, Bernard. Big Data: The 5 Vs Everyone Must Know, 2014. Disponível em: < https://www.linkedin.com/pulse/20140306073407-64875646-big-data-the-5-vs-everyone-must-know>. Acesso em: 16 de fevereiro de 2019.

MARR, Bernard. Here's Why Data Is Not The New Oil, 2018. Disponível em: < https://www.forbes.com/sites/bernardmarr/2018/03/05/heres-why-data-is-not-the-new-oil/#356d9ff93aa9>. Acesso em: 6 de março de 2019.

MICROSOFT. Microsoft R Open: The Enhanced R Distribution, 2018. Disponível em: <https://mran.microsoft.com/rro>. Acesso em: 6 de março de 2019.

NEO4J. Neo4j Graph Analytics - Your Path to Intelligent Applications, 2017. Disponível em: <https://go.neo4j.com/rs/710-RRC-335/images/Neo4j_Graph_Analytics.pdf>. Acesso em: 16 de fevereiro de 2019.

NIXON, Kamille. Sustainable Competitive Advantage: Creating Business Value through Data Relationships, 2018. Disponível em: <https://neo4j.com/whitepapers/sustainable-competitive-advantage-graph-databases/?ref=home>. Acesso em: 16 de fevereiro de 2019.

ORACLE. R Technologies from Oracle for Advanced Analytics, 2014. Disponível em: <https://www.oracle.com/technetwork/database/database-technologies/r/r-technologies/overview/index.html>. Acesso em: 6 de março de 2019.

PLATFORM. Graph. A Graph Platform Reveals and Persists Connections, 2019. Disponível em: < https://neo4j.com/product/>. Acesso em: 6 de março de 2019.

PRESS, Gil. Six Observations From A New Survey On The State Of Big Data Analytics, 2015. Disponível em: < https://www.forbes.com/sites/gilpress/2015/09/04/6-observations-from-a-new-survey-on-the-state-of-big-data-analytics/#7afee8127884>. Acesso em: 16 de fevereiro de 2019.

PROVOST, Foster; FAWCETT, Tom. Data Science for Business. USA, 2013.

RATHLE, Philip; SADOWKSI, Gorka. Fraud Detection: Discovering Connections with Graph Databases, 2017. Disponível em: <https://go.neo4j.com/rs/710-RRC-335/images/Neo4j_WP-Fraud-Detection-with-Graph-Databases.pdf>. Acesso em: 6 de março de 2019.

ROBINSON, Ian; WEBBER, Jim; EIFRE, Emil. Graph Databases. USA, 2015.

SAS. Big Data – O que é e qual a sua importância, 2019. Disponível em: < https://www.sas.com/pt_br/insights/big-data/what-is-big-data.html >. Acesso em: 16 de fevereiro de 2019.

TOMAR, Louisa et al. Big Data in the Public Sector: Selected Applications and Lessons Learned, 2016. Disponível em: < https://publications.iadb.org/en/big-data-public-sector-selected-applications-and-lessons-learned>. Acesso em: 6 de março de 2019.

VAN ERVEN, Gustavo C. Galvão; MDG-NoSQL: Modelo de Dados para Bancos NoSQL Baseados em Grafos. Tese (Mestrado Profissional em Computação Aplicada) – Departamento de Ciência da Computação da Universidade de Brasília. 2015.

Instituto Serzedello Corrêa

Missão

Aprimorar a Administração Pública em benefício da sociedade por meio do controle externo

Visão

Ser referência na promoção de uma Administração Pública efetiva, ética, ágil e responsável

TRIBUNAL DE CONTAS DA UNIÃO

tcu.gov.br/isc