DOI: 10.5748/18CONTECSI/PSE/KMG/6841

BIG DATA: KNOWLEDGE GENERATION THROUGH UNSTRUCTURED DATA

Lidia Gimenez Simao Macul Monteiro - UNIVERSIDADE DE SÃO PAULO - USP - Orcid: https://orcid.org/0000-0001-7260-1700

Fernando José Barbin Laurindo - UNIVERSITY OF SAO PAULO USP - Orcid: https://orcid.org/0000-0002-5924-3782

This article aims to improve learning on the process of unstructured data use. To develop the principal objective this study explores the relationship between big data (Volume, Velocity, Variety, Veracity and Value) and the DIKW pyramid (Data, Information, Knowledge, Wisdom). This article related two theoretical bases (Big Data and DIKW pyramid) to improve the learning on unstructured data. The reseach propositions are: The toughest aspects of unstructured data are data capture and data treatment; The knowledge generation using unstructured data depends of preliminary tacit knowledge. A multiple case study methodology was used to develop the article. A guide research was developed to investigate the research propositions in two companies. The article improve learning on the process of unstructured data. A multiple case study confirm the research propositions: The toughest aspects of unstructured data are data capture and data treatment; The knowledge generation using unstructured data depends of preliminary tacit knowledge. This article contribute improving the knowledge on unstrutured data use. Through two case study the reseach propositions were confirmed: Capture and data treatment are the toughest aspect and the knowledge Generation depends of preliminar tacit knowledge. This article explore two case study to identify through aspects of unstructured data use and a dependecy of tacit knowledge. The factors identified helps to learn about unstructured data use and identify factors that must be considered in knowledge management.

Keywords: Big Data, Knowledge Management, Strategy, Unstructured Data, DIKW

BIG DATA: GERAÇÃO DE CONHECIMENTO ATRAVÉS DE DADOS NÃO ESTRUTURADOS

O presente estudo tem como objetivo principal melhorar o entendimento sobre o processo de utilização de dados não estruturados para a gestão de conheciment0. Para isso o estudo utilizou duas bases teóricas: 5 V’s de Big Data e Pirâmide DICS (Dados, Informação, Conhecimento e Sabedoria). O estudo utiliza a relação entre duas bases teóricas (Big Data e Pirâmide DICS) para melhorar o entendimento sobre a geração de conhecimento através de dados não estruturados. As proposições de pesquisa analisam os aspectos mais complicados para utilizar estes dados e a dependência de conhecimento tácito. Como metodologia o estudo utilizou estudo de caso múltiplo aplicados em um bureau de dados e uma Startup. Foi utilizado um roteiro de pesquisa para direcionar a aplicação dos estudos de caso. Como resultado principal, o estudo foi capaz de melhorar o entendimento no uso de dados não estruturados e relacionar com as bases teóricas 5 V’s de Big Data e Pirâmide DICS, além de confirmar as duas proposições de pesquisa. O estudo contribui através da exploração de casos que utilizam dados não estruturados para a geração de conhecimento e através do relacionamento de duas bases teóricas, além disso o estudo foi capaz de confirmar as duas proposições de pesquisa. Como contribuição para gestão, através dos estudos de caso aplicados o estudo identificou os aspectos mais complicados na utilização de dados não estruturados e que devem ser considerados na gestão do conhecimento.

Palavras-chave: Big Data, Gestão do Conhecimento, Estratégia, Dados não estruturados, DICS

Big Data: Knowledge Generation through unstructured data

Big Data: Geração de Conhecimento através de dados não estruturados

ABSTRACT

This article aims to improve learning on the process of unstructured data use. To develop the principal objective this study explores the relationship between big data (Volume, Velocity, Variety, Veracity and Value) and the DIKW pyramid (Data, Information, Knowledge, Wisdom).

Different dimensions of big data affect unstructured data such as large volume, not stored in a traditional data base, diversified, accurate data and the ability to generate value. The use of unstructured data enables gain new information and knowledge.

Based on bibliographic search this study defines two research propositions: The toughest aspects of unstructured data are data capture and data treatment; The knowledge generation using unstructured data depends of preliminary tacit knowledge.

Using two case studies, this article confirms the research propositions. This study describes how the companies capture and treat unstructured data and the main difficulties to gain knowledge from this data.

The companies studied are from specific sector, therefore it is a limitation of study. Other companies need to be studied to explore new market sector and to clarify issues identified.

Keywords: Knowledge management, Big data, Strategy, unstructured data, DIKW.

RESUMO

O presente estudo tem como objetivo principal melhorar o entendimento sobre o processo de utilização de dados não estruturados para a geração de conhecimento. Para isso foi utilizada a base teórica dos 5 V’s de big data (Volume, Variedade, Velocidade, Veracidade e Valor) em conjunto com a pirâmide DICS (Data, Informação, Conhecimento, Sabedoria).

Dados não estruturados estão relacionados aos diferentes aspectos de big data, como o alto volume, não capazes de serem armazenados em um banco de dados tradicional com linhas e colunas, fluidez dos dados e em relação a confiabilidade nos dados e aos valores gerados.

A utilização de dados não utilizados anteriormente possibilita que uma empresa gere informações e conhecimentos diferenciados e diversificados que podem ser utilizados para a geração de estratégias.

Baseado no levantamento bibliográfico o estudo definiu duas proposições de pesquisas. A primeira proposição observa se a captura e tratamento de dados não estruturados é o aspecto mais complicado em projetos que utilizam esse tipo de dados, e a segunda proposição

observa se a geração de conhecimento através de dados não estruturados possui dependência prévia de conhecimento tácito.

Através do estudo de caso múltiplo as duas proposições de pesquisa foram confirmadas. O presente estudo descreve como as empresas capturam e tratam dados não estruturados e as principais dificuldades para transformar estes dados em conhecimento.

O estudo tem como fator limitante a aplicação de dois estudos de casos aplicados em setores específicos. Futuros estudos deverão ampliar o volume e setores de empresas estudadas, assim como clarificar pontos destacados no presente estudo.

PALAVRAS-CHAVE: Gestão do Conhecimento, Big Data, Estratégia, Dado não estruturados, DICS.

I. INTRODUÇÃO

Cada vez mais as organizações são capazes de utilizar dados e informações para a geração de conhecimento. As evoluções tecnológicas se tornaram facilitadores para a geração de conhecimento através de dados, uma vez que são capazes de tornar o processo de captura, tratamento e análise de dados em informações de forma mais eficiente e eficaz.

A principal razão que tem levado as empresas a utilizar o recurso de dados é a possibilidade de geração de conhecimento sobre a própria empresa, sobre os clientes e sobre os competidores (Zelenka & Podares, 2021). Dessa forma, o conhecimento gerado por meio da informação tem capacidade de proporcionar para as empresas a compreensão diante de um contexto específico, além de auxiliar na tomada de decisão estratégica.

Para Davenport (1998) o conhecimento é valioso porque fornece um significado e uma interpretação das informações diante de um contexto.

Seguindo a lógica da pirâmide DICS – Dados, Informações, Conhecimento e Sabedoria (Faucher, Everett & Lawson, 2008), através da captação, estruturação e transformação dos dados, as empresas são capazes de obter informações que levem ao conhecimento e a sabedoria.

Informação é um conjunto de dados devidamente tratados de forma a serem providos de significado, bem como organizados e classificados para alguma finalidade e úteis para as pessoas em processo de tomada de decisão (Laurindo, 2008).

Por sua vez, conhecimento é um conjunto de ferramentas conceituais e categorias usadas pelas pessoas para criarem, colecionarem, armazenarem ou compartilharem informação (Laurindo, 2008).

A TI (tecnologia da informação) se torna imprescindível para a captação, tratamento e armazenamento de dados, assim como para a geração de informações. Informação está relacionada aos dados armazenados nos sistemas de informação e no conhecimento gerado a partir da análise da informação gerada pelos dados (Davenport, 1998).

A captação e utilização de diferentes dados possibilitam a geração de informações e conhecimentos diversificados que podem ser utilizados pelas empresas para a geração de estratégias e para a geração de vantagem competitiva.

Cada vez mais as evoluções em TI permitem a exploração de novos dados. Dados volumosos, não estruturados ou dinâmicos, ou seja, iniciativas em relação a *big data* são cada vez mais exploradas pela TI.

Big data pode ser definido através de 5V's: Volume, Variedade, Velocidade, Veracidade e Valor (Cheng, Zhang & Qin, 2016). Note-se que as três primeiras características é que de fato são específicas de *Big Data*, uma vez que as duas últimas são exigidas de quaisquer aplicações de TI.

Big data é um termo utilizado para dados volumosos demais para serem armazenados em um único servidor, dados não estruturados para se adequar a um banco de dados organizados em linhas e colunas, ou dados fluídos demais para serem armazenados em um *data warehouse* estático. Embora o termo enfatize o tamanho, o aspecto mais complicado de *big data* envolve sua falta de estrutura (Davenport, 2014).

O presente estudo tem como objetivo principal melhorar o entendimento sobre o processo de utilização de dados não estruturados para a geração de conhecimento.

II. REVISÃO TEÓRICA

II.1 GESTÃO DO CONHECIMENTO

A literatura sobre gestão do conhecimento está repleta de definições sobre dados, informação, conhecimento e sabedoria (Faucher *et al*., 2008). Usualmente se considera que conhecimento é obtido através de dados e informações.

Davenport (2000) define dados, informação e conhecimento como:

- Dados são registros de transações, fatos e eventos.
- Informação está relacionado à mensagem transmitida e é capaz de mudar a percepção sobre algo.
- Conhecimento está relacionado com a experiência, valores, informação contextual e especialidade que proporcionam a evolução e incorporação de novas experiências e informação.

Sabedoria é considerado como o conhecimento que foi processado de forma significativa, e está relacionado ao senso crítico de utilização do conhecimento de forma construtiva e discernimento para a criação de novas ideias (Faucher *et al*., 2008). A sabedoria é uma evolução do conhecimento (Ackoff, 1989).

O conhecimento pode ser separado em dois tipos: Conhecimento tácito e conhecimento explícito (Nonaka & Takeuchi, 2001).

Conhecimento explícito pode ser expressado em uma linguagem formal e sistêmica, e pode ser facilmente compartilhada, enquanto que o conhecimento tácito é utilizado com maior dificuldade porque é pessoal e subjetivo (Faucher *et al*., 2008). O conhecimento tácito está relacionado às experiências e conhecimentos prévios.

A percepção natural é que dados e informação são explícitos, enquanto que conhecimento e sabedoria são tácitos, porém os aspectos do tácito e do explícito estão presentes desde o nível de dados até o nível de sabedoria (Faucher *et al*., 2008).

A Figura 1 mostra a relação entre tácito e explícito ao longo da hierarquia DICS (dado – informação – conhecimento - sabedoria).

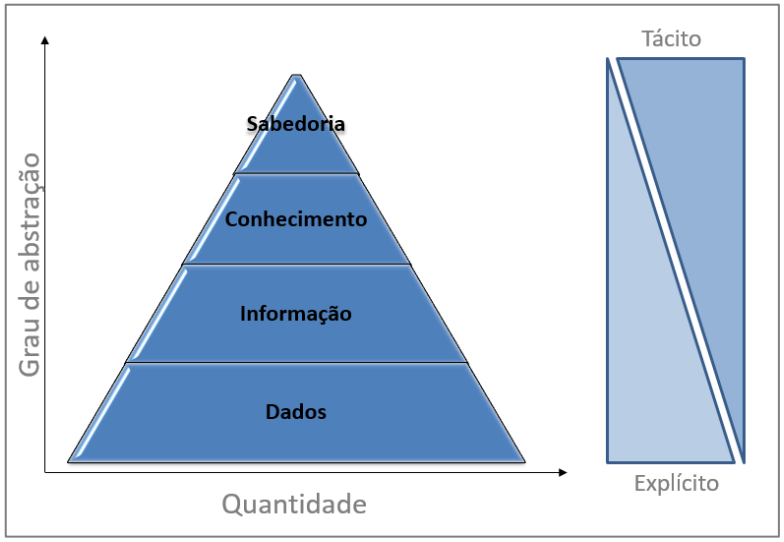

Figura 1 – Hierarquia DICS
Fonte: Faucher; Everett; Lawson (2008)

A pirâmide DICS apresenta a hierarquia da geração do conhecimento. O primeiro nível, relativo aos dados, é caracterizado por alto volume e baixo grau de abstração e com o conhecimento explícito maior que o conhecimento tácito. Os níveis intermediários são caracterizados sucessivamente por informação e conhecimento, sendo que informação existe em maior quantidade e menor grau de abstração do que conhecimento. No nível de informação o explícito é maior do que o tácito, enquanto que no nível de conhecimento o tácito é maior do que o explícito.

O último nível da pirâmide DICS seria aquele com baixo volume e alto grau de abstração, e com conhecimento tácito maior do que o conhecimento explícito.

A pirâmide DICS mostra que ao longo da hierarquia o grau de abstração e a predominância do conhecimento tácito aumentam, enquanto que a quantidade e conhecimento explícito diminuem, ou seja, quanto maior for a abstração, maior é a dependência da experiência e do conhecimento prévio dos envolvidos.

A transformação de dado em informação é bem realizada pela TI. Entretanto a transformação de informação em conhecimento, por envolver aspectos humanos e sociais, é bem menos suportada pela TI, pois é um processo que não pode ser automatizado (pelo menos não inteiramente) (Laurindo, 2008).

Aspectos humanos estão relacionados com a experiência e contextualização dos envolvidos, e podem influenciar na geração do conhecimento.

As estratégias baseadas no conhecimento gerado pelas empresas estão diretamente relacionadas ao processo de geração de conhecimento adotado pelas empresas (Zack, 1999).

A capacidade de uma empresa de transformar dados em informação e em conhecimento está relacionado com os conhecimentos tácitos e explícitos existentes, porém tem potencial diferenciador para o direcionamento de estratégias e para melhorar a capacidade de agir.

II.2 *BIG DATA*

Big data é um termo genérico para dados que não podem ser contidos nos repositórios usuais; refere-se a dados volumosos demais para serem armazenados em um único servidor; não estruturados demais para se adequar a um banco de dados organizados em linhas e colunas; ou fluídos demais para serem armazenados em um *data warehouse* estático. Embora o termo enfatize o tamanho, o aspecto mais complicado de *big data* envolve sua falta de estrutura (Davenport, 2014).

Iniciativas em *big data* são cada vez mais adotadas pelas corporações e podem influenciar a forma de geração do conhecimento, uma vez que *big data* é capaz de proporcionar a exploração de dados com maior diversificação. A exploração de novos dados ou dados mais detalhados permite que as empresas possuam novas informações que podem ser utilizadas para de fornecer conhecimento.

Quando é possível obter dados mais detalhados e fazer uma análise mais sistemática desses dados, o resultado provavelmente será a obtenção de decisões estratégicas (Davenport, 2014).

Joey Fitts (CEO da Matters Corp) explica que historicamente as informações coletadas sobre o mercado e o setor indicavam quem formava o setor, mas agora também é possível identificar o que fazem. Fatores de mercado até então ocultos agora são visíveis, possibilitando a análise de tendências, *benchmarking*, segmentação, modelagem e recomendações. É um conjunto muito mais amplo de dados, em uma escala muito maior e mais em tempo real. Agora as empresas podem liderar ao invés de apenas reagir (Davenport, 2014).

O conhecimento obtido através de novas informações pode levar as empresas a diferentes decisões estratégicas, por exemplo, a informação de fluxo de pessoas dentro de uma loja, captada através de dados de câmeras de monitoramento, pode influenciar na decisão da disposição dos produtos dentro dessa loja. Dados de voz de clientes podem ser utilizados para auxiliar no nível de satisfação e fidelização dos consumidores. A exploração de dados tem a capacidade de gerar novos conhecimentos para uma empresa.

Big data é definido através de 5 V's (Davenport, 2014) (Cheng *et al.*, 2016):

- Variedade: diferentes tipos de dados e origens, o que gera dificuldade de adequação a um banco de dados tradicional;
- Volumosos: grande quantidade de dados, o que gera dificuldade de armazenamento;
- Velocidade: fluídos demais para serem armazenados em um *data warehouse* estático.
- Veracidade: confiabilidade nos dados. Dados inconsistentes e incompletos dificultam ou não permitem seu uso.
- Valor: retorno gerado em relação a valor, percepções e benefícios gerados.

Para Davenport (2014) veracidade e valor são fatores importantes para *big data* e devem ser levados em consideração, porém isoladamente não representam *big data*, enquanto que os três primeiros componentes (volume, variedade e velocidade) existindo isoladamente ou em conjunto podem determinar dados de *big data*. Dessa forma, *big data* pode ser composto por um ou mais dos três primeiros componentes definidos de *big data*, porém veracidade e valor devem ser considerados em conjunto. Além disso, veracidade e valor são características que se espera de todas as aplicações de TI, não especificamente de aplicações de *Big Data* (que é um tipo de aplicação de TI).

Big data analytics é o termo utilizado para a análise de dados *big data. Big data analytics* é definido como a extração de informação e conhecimento de uma grande quantidade de dados, velozes e/ou voláteis e que ocorre através da mineração destes dados e do reconhecimento de padrões (Cheng, *et al*., 2016).

Big data analytics fornece informações que podem ser utilizadas para o direcionamento de decisões, podendo ser gerada em tempo real em todos os dados através do uso de algoritmos automatizados (BRATA, 2014).

Um dos aspectos de *Big Data Analytics* envolve a falta de estruturação de dados e a necessidade de trabalhar esses dados para que para que seja possível a extração de conhecimento, o que fornece para as empresas a possibilidade de obter um melhor direcionamento estratégico.

Big Data cria oportunidade de descobrir e obter conhecimento baseados em informação melhorando o negócio principalmente em predição e tomada de decisão, porém um grande desafio para as empresas é a capacidade de adaptação às novas tendências e criar novos conhecimentos para ativar vantagens competitivas sustentáveis (Sumbal, Tsui & See-To, 2017).

III. METODOLOGIA

Para o desenvolvimento do presente estudo foram analisados os 5 V's de *big data* considerando a forma de captura, estruturação e armazenamento de dados não estruturados, mas que podem também apresentar volatilidade e alto volume. Outros fatores como veracidade e valor gerado também foram discutidos, assim como a capacidade de agir e direcionamento de estratégias.

Para a investigação do objetivo principal de estudo foram definidos objetivos específicos:

- Compreender o processo de captura, processamento, armazenamento e organização de dados;
- Compreender o processo de transformação de dados em informação;
- Compreender o processo de geração do conhecimento e sabedoria;
- Compreender como as empresas realizam a gestão do conhecimento para a geração de estratégias.

Para o desenvolvimento do estudo foi utilizada a metodologia de Estudo de caso múltiplo com a utilização de um roteiro de pesquisa.

Os critérios de seleção das empresas para o desenvolvimento do estudo de caso foram os seguintes:

- Empresas que desenvolvem iniciativas em *big data* utilizando dados não estruturados;
- Empresas que geram informação e conhecimento através da utilização destes dados não estruturados;
- Empresas que utilizam essas informações e conhecimentos para a geração de estratégias;
- Empresas que possibilitem acesso a pelo menos um entrevistado que tenha um cargo gerencial na empresa, e que deve estar envolvido nos processos e estratégias utilizados.

A partir do objetivo de pesquisa e do levantamento bibliográfico realizado, o presente estudo analisará as seguintes proposições de pesquisa:

> P1: A captura e tratamento de dados não estruturados é o aspecto mais complicado para a geração de conhecimento em projetos que utilizam esse tipo de dados (Davenport, 2014) (Sumbal *et al.*, 2017).
>
> P2: A geração de conhecimento através de dados não estruturados possui dependência prévia de conhecimento tácito (Faucher *et al*., 2008) (Zack, 1999).

IV. ESTUDO DE CASO

Para a aplicação do estudo de caso foram selecionadas empresas dentro dos critérios:

- Empresas que desenvolvem iniciativas em big data utilizando dados não estruturados;
- Empresas que geram informação e conhecimento através da utilização destes dados não estruturados;
- Empresas que utilizam essas informações e conhecimentos para a geração de estratégias;
- O entrevistado deve ter um cargo gerencial na empresa, e deve estar envolvido nos processos e estratégias utilizados.

Conforme os critérios de pesquisa definidos, duas empresas foram selecionadas e estudadas: um *bureau* de informações e uma *startup*.

A primeira empresa estudada é um *bureau* de informações. Essa empresa tem o foco em soluções de crédito, marketing, certificação e consulta de dados. Para o desenvolvimento deste estudo um único produto da empresa foi selecionado: *Voice Analytics*.

O produto estudado tem como objetivo a identificação de padrões de conversas telefônicas. As gravações telefônicas são transcritas, e padrões de frases são identificados com um propósito específico.

Duas aplicações foram descritas pelo entrevistado. A primeira é a identificação de fraude de produtos vendidos através de ligação telefônica, como por exemplo, a identificação de fraude nas vendas de chips de celular onde mesmo com o cliente negando o interesse em comprar o chip, o atendente faz a marcação de que a venda foi aceita.

Segundo o entrevistado, esse problema ocorre devido principalmente a meta de vendas determinada pela empresa e a bonificação dos vendedores, ou seja, para atingir a meta e obter a bonificação os vendedores concluem a venda mesmo contra a vontade do consumidor. Posteriormente o cliente recebe erroneamente o produto e precisa cancelar. A empresa além de ter o prejuízo da venda não executada, da bonificação dada ao funcionário (a bonificação pode ocorrer antes do cancelamento) e de todos os custos envolvidos com envio e cancelamento do produto, existe principalmente um dano à imagem da empresa perante o consumidor, que pode ser ampliada quando divulgada nas redes sociais e, além disso, existe a possibilidade de o consumidor entrar com ação cível ou com reclamação em órgão regulador.

Outra utilização deste produto é com a finalidade de segmentação de consumidores, por exemplo, em uma empresa de cobrança, onde a empresa realiza ligações com o intuito de

quitação da dívida, e através da transcrição da conversa e da identificação de padrões é possível segmentar os consumidores devedores em grupos (possibilidade de acordo, não vai quitar a dívida, não consegue quitar, desempregado, etc.). Através da segmentação é possível definir diferentes estratégias de atuação perante os consumidores.

O estudo de caso com o bureau foi aplicado com o gerente de produtos de *big data*. O entrevistado conhece bem o produto e participa da implantação deste produto em suas diferentes aplicabilidades.

A segunda empresa estudada é uma *startup* que desenvolveu um produto com foco em identificação do comportamento do consumidor em relação a trajetória realizada em locais como *shoppings*, eventos, parques com o intuito de identificar padrões de comportamento, por exemplo, maior concentração em lojas que vendem determinado produto no *shopping*, ou maior tempo observando determinado estande em um evento.

Os consumidores precisam inicialmente concordar em ter seu rosto escaneado para que a identificação facial exata do consumidor possa ocorrer, portanto é necessário o aceite prévio dos mesmos. Através das imagens de câmeras é possível identificar a trajetória exata de cada consumidor. O intuito desse produto é poder identificar padrões de comportamento e poder melhorar a estratégia de distribuição das lojas dentro do shopping, estandes em eventos ou atrações de um parque. Embora esse produto já tenha sido desenvolvido e testado em pequena escala, ainda não foi efetivamente implantado. O estudo de caso da startup foi aplicado com um dos desenvolvedores do produto.

A aplicação do estudo de caso foi dividida em três partes:

1. Estruturação dos dados. Compreendendo a forma de captação e estruturação dos dados, e compreendendo as dificuldades e limitantes dos dados big data e suas características.
2. Obtenção de informações. Compreendendo como os dados são transformados em informação acionável.
3. Conhecimento e sabedoria. Compreendendo como o conhecimento e sabedoria são obtidos, e como a capacidade de agir e definição de estratégias geradas através de dados não estruturados.

V. RESULTADOS

A análise dos estudos de caso realizados permitiu que as proposições de pesquisas fossem investigadas.

> P1: A captura e tratamento de dados não estruturados é o aspecto mais complicado para a geração de conhecimento em projetos que utilizam esse tipo de dados.
>
> P2: A geração de conhecimento através de dados não estruturados possui dependência prévia de conhecimento tácito.

Empresa 1: *Bureau.*

O produto de *Voice Analytics* desenvolvido pela empresa utiliza a gravação das chamadas telefônicas para obtenção dos dados. As empresas avisam previamente seus clientes que a ligação está sendo gravada, pois é necessário ter esse aceite para a captura destes dados.

Para a empresa o principal limitante do produto é a transcrição das vozes para texto, ou seja, o tratamento de dados não estruturados. A transcrição é realizada por uma empresa parceira, de forma que a empresa estudada envia os arquivos de voz para a empresa parceira, e recebe como retorno um arquivo com a transcrição dessas vozes no formato de texto. Segundo o entrevistado, cerca de 80% das transcrições são realizadas com sucesso, porém ainda há a perda de 20% que foram identificadas como principal fator de insucesso a qualidade das gravações das chamadas.

Com o intuito de aumentar esse percentual de transcrição, o Bureau solicitou a seus clientes que aumentassem a qualidade dessas gravações. Para as empresas que seguiram a recomendação, o percentual de retorno aumentou, porém consequentemente foi necessário um aumento no espaço de armazenamento destas ligações, o que para outras empresas clientes do bureau se apresentou como fator limitante.

Após a transcrição das vozes para texto é necessário a identificação de padrões de linguagem. Algoritmos de identificação de padrões são desenvolvidos de duas formas principais: A primeira é a identificação ontológica, baseada na linguagem natural, ou seja, a linguagem baseada em um dicionário de dados, onde existe a identificação de palavras e conjunto de palavras relevantes ao negócio, por exemplo, no caso de identificação de fraude em venda de chip de celulares, a identificação de frases como “não tenho interesse” ou “não quero” identificam o não interesse do consumidor.

A segunda forma de identificação de padrões ocorre através de *machine learning*, e se baseia no contexto da conversa, o que permite a identificação se o contexto é bom ou ruim, ou seja, é possível perceber indícios de ironia dentro da conversa, além de ser possível melhorar o algoritmo de identificação de padrões conforme mais conversas são analisadas. Um exemplo citado pelo entrevistado é o uso da palavra “querida” que mesmo sendo uma palavra positiva muitas vezes é utilizada em tom de ironia.

O processo de identificação de padrões é um processo robusto e depende do armazenamento do texto completo da ligação. Para solucionar o problema de armazenamento a empresa utiliza armazenamento em nuvem, pois dessa forma somente o espaço necessário é utilizado durante o processo de identificação de padrões, e a empresa só tem o custo do espaço utilizado.

A identificação de padrões em português brasileiro ainda é um limitante. O algoritmo de transcrição de voz em texto consegue ser aprimorado conforme mais transcrições são realizadas e neste caso, além de haver grandes diferenças regionais de linguagem que influenciam diretamente em aspectos como ironia, somente os dados do Brasil podem ser considerados para esse aprimoramento, o que afeta inclusive a identificação ontológica devido aos diferentes sotaques regionais.

O processo de captura, transcrição e identificação de padrões de linguagem é o processo de transformação de dados em informação, ou seja, é onde ocorre a análise de dados *big data* (ou *big data analytics*). Essa etapa do processo foi definida pelo entrevistado como a parte mais robusta do processo e que depende de um conhecimento prévio dos envolvidos, pois é necessário que previamente já se tenha conhecimento das informações desejadas.

A informação obtida varia conforme a empresa na qual ocorreu a aplicação. Por exemplo, para uma empresa de telefonia celular o retorno é separação em dois grupos de solicitou ou não um chip. Posteriormente essa informação é confrontada com a informação de venda. Dessa forma a empresa consegue gerar informações e ter o conhecimento para trabalhar seu foco estratégico: identificação de fraudes. Essa identificação auxilia a empresa uma vez que

impede o envio de produtos indesejados aos clientes e além de não prejudicar a imagem da empresa facilita a identificação de funcionários fraudulentos.

Já para uma empresa de cobrança o retorno é a segmentação em três grupos: irá realizar acordo, não irá realizar acordo e impossibilitado de fazer acordo, com a justificativa (ex: desempregado, etc.). Essa segmentação em grupos ocorre através de modelos estatísticos que utilizam padrões de linguagens identificadas e definem a propensão à quitação da dívida. Dessa forma a empresa pode priorizar seus contatos e pode focar seus esforços de cobrança no público mais propenso a quitar a dívida.

Para as duas formas de identificação de padrões de conversa é necessário um conhecimento prévio dos desenvolvedores sobre o assunto a ser tratado, e sobre as estratégias a serem adotadas. O padrão de conversas é diferente conforme o segmento de mercado, portanto o conhecimento prévio de frases e comportamentos relacionados ao segmento é um fator importante para o desenvolvimento dos algoritmos de reconhecimento de padrões.

Dessa forma, para essa aplicação de big data o conhecimento tácito se apresentou como um conhecimento que influencia no processo de obtenção de informação, conhecimento e sabedoria.

Empresa 2: *Startup*

O produto desenvolvido pela empresa 2 é um produto utilizado pontualmente devido ao seu principal limitante de precisar digitalizar previamente o rosto das pessoas e ter o aceite dos mesmos, porém esse é um produto que utiliza os conceitos de *big data* para a geração do conhecimento.

O produto tem como objetivo principal determinar o padrão de fluxo das pessoas, por exemplo dentro de um *shopping*, sendo possível mensurar o caminho percorrido entre as lojas, considerando os diferentes andares, além de ser possível mensurar o tempo entre as transições considerando cada cliente único. Esse produto pode ser também utilizado em eventos ou parques com o objetivo de identificar trajetórias e concentrações.

Inicialmente, como dito, é necessário digitalizar a face dos consumidores para que a identificação possa ocorrer e também é necessário obter uma aceitação para esse rastreamento. Devido às leis de regularização, proteção e sigilo de dados somente podem ser rastreados e armazenados dados de clientes que derem um aceite.

Para o entrevistado esse ainda é um problema que impacta diretamente na viabilização do produto, pois as pessoas não se sentem confortáveis em serem rastreadas, ou podem ter atitudes diferenciadas, uma vez que sabem que estão sendo observadas. A principal forma de incentivo ao cadastramento por parte dos clientes é o fornecimento de descontos e brindes. Outra limitação do produto é a existência de grande semelhança entre pessoas, porém esse erro já analisado e classificado pelo entrevistado como pequeno.

A identificação facial ocorre através de câmeras existentes nos ambientes e a identificação do cliente único se baseia na simetria da face, ou seja, na comparação entre lado direito e esquerdo da face, distância entre os olhos, etc.

Assim como no produto desenvolvido pela empresa 1, a transcrição é a parte mais complexa e de maior custo do produto. As imagens devem ser analisadas em ordem cronológica e além da trajetória percorrida fornecem o tempo entre localizações e o tempo em determinadas localizações.

O mapeamento de trajetórias permite a observação de diferentes padrões de comportamento, o que possibilita a determinação de diferentes grupos de comportamento nos locais analisados. Por exemplo, um segmento transita pelo *shopping* buscando lojas de eletrônicos e posteriormente vai para a praça de alimentação, outro segmento busca por artigos esportivos e calçados, já em eventos ou parques é possível observar se existe um fluxo mais seguido e se as pessoas se mantêm mais tempo em algum estande ou atração. Essa informação auxilia em estratégias voltadas para definição de quais atrações ou produtos são mais interessantes para o público.

Através da identificação de comportamentos dos consumidores diferentes estratégias podem ser adotadas, por exemplo é possível modificar a distribuição de segmentos de lojas dentro do *shopping* buscando estimular o trafego entre dois ou mais tipos de lojas para um público que busca esses determinados produtos, ou é possível mudar a localização de estande e atrações para estimular o trafego entre dois pontos de maior concentração.

Essa análise auxilia na identificação de diferentes padrões de comportamento do público, ou seja, é possível mensurar e classificar diferentes perfis de comportamento.

O conhecimento prévio das informações de interesse é necessário, e impacta no desenvolvimento e aplicação da solução desenvolvida. Para que os dados sejam tratados e gerem as informações de interesse é necessário compreender previamente quais são as informações de interesse. Para a aplicação do produto em *shoppings*, além de identificação de trajetória e tempo observando vitrines de lojas foi identificado que existe diferença de comportamento conforme horário e quantidade de pessoas no grupo, o que auxilia em decisões focadas na atratividade do público consumidor.

Assim como no estudo de caso aplicado na empresa 1, o entrevistado da empresa 2 também definiu que o conhecimento tácito é necessário desde o início do processo, assim como o conhecimento prévio de direcionamento de estratégias.

Através dos dois estudos de caso aplicados foi possível compreender como as empresas pesquisadas capturam e estruturam dados em contexto de *big data*, e como estes dados são transformados em informação. Além disso, o estudo identificou que a transformação das informações em conhecimento acionável para a geração de estratégias depende de um direcionamento prévio sobre as estratégias a serem adotadas.

Para o tratamento do alto volume de dados não estruturados foram utilizados recursos como o armazenamento em nuvem e somente dados previamente selecionados são armazenados, ou somente as informações previamente consideradas relevantes são armazenadas. Os dados não estruturados possuem alto volume, o que também se apresentou como fator de impacto para a transformação destes dados em informação.

Nos estudos de caso aplicados a variedade de origens não foi identificada, porém a dificuldade de adequação a um banco de dados tradicional foi identificada. Para solucionar este problema as empresas precisam previamente saber identificar quais são as informações relevantes que levam ao conhecimento desejado.

Dentro dos casos estudados a fluidez dos dados não é tratada em tempo real. A velocidade de tratamento de dados é limitada principalmente devido a limitação tecnológica de transcrição da informação.

A veracidade dos dados é analisada durante a implantação do produto. Segundo os entrevistados é necessária uma verificação da qualidade e da confiabilidade dos dados no momento da implantação. No estudo de caso 1 a confiabilidade está diretamente relacionada

ao conhecimento prévio do segmento de mercado, o que impacta diretamente no desenvolvimento dos algoritmos de identificação de padrões implantados.

Em relação ao valor, ao retorno e aos benefícios gerados é necessário que o conhecimento desejado seja previamente definido uma vez que a estruturação dos dados ocorre de forma focada no conhecimento almejado. Em relação ao retorno financeiro, no primeiro estudo de caso é possível mensurar a economia financeira gerada, porém no segundo estudo de caso, embora exista um retorno financeiro esperado o produto não foi utilizado em grande escala e não se pode confirmar esse retorno.

A geração de conhecimento e sabedoria devem ser previamente estimados devido ao alto esforço e custo existentes, ou seja, existe uma dependência do conhecimento prévio sobre o direcionamento estratégico a ser adotado.

A Tabela 1 apresenta uma síntese da análise dos casos estudados e como os 5 V's de *Big Data* se manifestam em cada empresa.

	Bureau	***Startup***
Volume	O alto volume de dados de voz gerou a necessidade de utilização de recurso diferenciado: Armazenamento em Nuvem	O alto volume de dados de imagem influencia diretamente em conhecer previamente as informações que se deseja obter.
Velocidade	Os dados são gerados em alta velocidade, porém a necessidade de transcrição utilizando uma empresa terceira impacta diretamente na velocidade com que as informações são obtidas.	Embora os dados sejam gerados com alta velocidade, dado a necessidade de transcrição não é possível obter informações em tempo real.
Variedade	Existe necessidade de transformar os dados não estruturados em um banco de dados tradicional, para isso o conhecimento desejado deve ser previamente conhecido além da necessidade de se conhecer o negócio para que a análise seja desenvolvida corretamente.	Os dados não estruturados são transformados em estruturados, porém é necessário ter conhecimento prévio do conhecimento que se deseja obter.
Veracidade	Está diretamente relacionada ao conhecimento prévio do segmento de mercado, e que impacta diretamente no desenvolvimento dos algoritmos de identificação de padrões implantados	É necessária uma verificação da qualidade e da confiabilidade dos dados na captura e principalmente na transcrição.
Valor	O conhecimento gerado precisa ser previamente identificado para que a transcrição e análise ocorra direcionada para o conhecimento	O conhecimento almejado foi obtido, porém o retorno financeiro pode ser pouco observado devido a

	desejado. Em relação a valor financeiro, as empresas que utilizaram o produto obtiveram melhores resultados.	sua utilização ter ocorrido de forma pontual.

Quadro 1 – Síntese dos casos
Fonte: Desenvolvido pelos autores

Dessa forma as proposições de pesquisa foram analisadas:

> P1: A captura e tratamento de dados não estruturados é o aspecto mais complicado para a geração de conhecimento em projetos que utilizam esse tipo de dados.
>
> A P1 foi confirmada nos estudos de casos aplicados. Os dois entrevistados confirmaram que o aspecto mais complicado é a falta de estrutura dos dados e como tratar os dados não estruturados, mas também em relação ao seu alto volume pois é necessário utilizar recursos (como nuvem) que possibilite o armazenamento destes dados mesmo que temporariamente, o que pode se tornar fator limitante uma vez que algumas empresas podem não querer investir nesses recursos.
>
> P2: A geração de conhecimento através de dados não estruturados possui dependência prévia de conhecimento tácito.
>
> A segunda proposição de pesquisa foi confirmada. Para o tratamento de dados não estruturados as empresas dependem previamente de um conhecimento tácito sobre o assunto investigado e principalmente dependem de já terem um direcionamento da informação a ser gerada. Dados não estruturados são trabalhados e dados estruturados são armazenados, por exemplo, suspeita de fraude ou trajetória em um ambiente.

VI. CONCLUSÃO

O conhecimento é obtido através de dados e das informações. A obtenção de conhecimento a partir de dados não estruturados de *big data* exige um grande esforço no primeiro nível da hierarquia, ou seja, no nível dos dados. Para a estruturação e uso dos dados é necessária uma expectativa prévia do conhecimento a ser obtido, e é necessário ter um conhecimento prévio das possíveis estratégias a serem adotadas.

Em *big data* o processo de geração de conhecimento segue a hierarquia DICS, através dos dados *big data* e de *big data analytics*, porém a dependência do conhecimento tácito se mostrou com alta importância devido a necessidade de estruturação destes dados para que possam ser transformados em informação e, em seguida, em conhecimento e em sabedoria.

A experiência prévia, valores, informação contextual e especialidade dos envolvidos impactam diretamente no conhecimento a ser gerado, e consequentemente influenciam no direcionamento estratégico a ser adotado devido ao alto grau de abstração dos dados.

O presente estudo contribui através da investigação da utilização de *big data* e *big data analytics* para a geração de conhecimento. Para dados não estruturados e a transformação destes dados em informação existe um grande esforço, ou seja, foi identificado pelos estudos

de casos aplicados que na base da hierarquia DICS há um grande esforço e se torna o processo mais complicado para a geração de conhecimento.

Conforme destacado na literatura, as grandes dificuldades de utilização de *big data* estão focadas principalmente em relação à falta de estruturação destes dados, à dificuldade de estruturação desses dados e à necessidade de experiência prévia dos envolvidos que no momento de sua implantação não podem ser automatizados inteiramente.

As evoluções tecnológicas proporcionam atualmente a utilização desses dados, e possivelmente outras evoluções tecnológicas futuras farão com que essas e outras dificuldades sejam tratadas com maior facilidade e com menor custo. Todavia, essas evoluções relacionadas a *big data* deverão ser estudadas conforme ocorrerem as evoluções tecnológicas.

O estudo possui como fator limitante a aplicação de dois estudos de caso que possuem produtos desenvolvidos para setores específicos. Outros setores devem ser analisados como evolução da pesquisa. Além disso, fatores apontados como limitantes (por exemplo, espaço de armazenamento) podem a longo prazo, dependendo das eventuais evoluções tecnológicas, deixarem de ser fatores limitantes.

Futuros estudos deverão ampliar o número de empresas estudadas, contemplando outros setores de atividades e também visando mais clarificação dos pontos destacados no presente trabalho.

BIBLIOGRAFIA

Ackoff, R.L. (1989). From data to wisdom. Journal of Applied Systems Analysis, 16, 3-9.

Brata, S. (2014). Big data analytics and its reflections on DIKW hierarchy. Review of management, 4 (1/2), 5-17.

Cheng, S., Zhang, Q., & Qin, Q. (2016). Big data analytics with swarm intelligence. Industrial Management & Data Systems, 116, 4.

Davenport, T. H. (1998). Ecologia da informação: por que só a tecnologia não basta para o sucesso na era da informação. São Paulo, Futura.

_. (2014). Big data at work: dispelling the myths, uncovering the opportunities. Havard Business Review.

_, Prusak, L. (2000). Working Knowledge: How Organizations Manage What They Know. Havard Business School Press. Boston, Massachussets.

Faucher, J.B., Everett, A. M. & Lawson, R. (2008). Reconstituting knowledge management. Journal of Knowledge Management, 12(3), 3–16.

Kelenka, M., Podaras, A. (2021). Increasing the effectivity of Business Intelligence Tools via Amplified Data Knowledge. Studies in Informatics and Control, 30(2), 67-77.

Laurindo, F. J. B. (2008). Tecnologia da informação: Planejamento e gestão de estratégias. São Paulo: Atlas.

Nonaka, I., & Takeuchi, H. (1995). The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation, Oxford University Press, New York, NY.

Sumbal, M. S., Tsui, E., & See-To, W.K. (2017). Interrelationship between big data and knowledge management: an exploratory study in the oil and gas sector. Journal of Knowledge Management, 21(1), 180-196.

Zack, M.H. (1999). Managing codified knowledge, Sloan Management Review, 40(4), 45-58.