ISSN: 2317-8302

DESCOBERTA DE CONHECIMENTO EMPRESARIAL COM DADOS ORIUNDOS DE DISPOSITIVOS IOT: UMA PROPOSTA PARA O DATA LAKE

BUSINESS KNOWLEDGE DISCOVERY WITH IOT DEVICES GENERATED DATA: A DATA LAKE PERSPECTIVE

MATEUS PINTO DA LUZ LIGOCKI VIEIRA
UNIVERSIDADE DE SÃO PAULO - ESCOLA POLITÉCNICA

DESCOBERTA DE CONHECIMENTO EMPRESARIAL COM DADOS ORIUNDOS DE DISPOSITIVOS IOT: UMA PROPOSTA PARA O DATA LAKE

Objetivo do estudo
Este estudo teve como objetivo analisar, no âmbito teórico, a possibilidade de se utilizar uma estrutura de Data Lake, para realizar a descoberta de conhecimento em dados gerados por dispositivos IoT; visando uma aplicação no contexto empresarial.

Relevância/originalidade
Este artigo tem relevância em razão de não ter sido encontrado trabalhos prévios que correlacionassem os conceitos nem utilização de dispositivos IoT, Data Lake e descoberta de conhecimento, o que dificulta a aplicação e estudo da descoberta do conhecimento neste contexto.

Metodologia/abordagem
Utilizou-se uma abordagem teórica para revisitar os trabalhos anteriores que estudaram e discutiram esses conceitos, isolada ou binariamente, e depois correlacionar os temas e analisar a viabilidade de descobrir conhecimento, utilizando um Data Lake, com dados de dispositivos IoT.

Principais resultados
Dentre os principais resultados, há que os dados IoT são heterogêneos, produzidos em alta velocidade e volume, a estrutura de Data Lake permite otimizar o armazenamento e recuperação dos dados de forma a gerar insumos flexíveis para a descoberta do conhecimento.

Contribuições teóricas/metodológicas
Este artigo fornece como contribuição teórica o estudo, em conjunto, do conceito de Data Lake, dispositivos IoT e descoberta do conhecimento, analisando as possibilidades e dificuldades, o que ensejará, principalmente, estudos práticos sobre essa abordagem conjunta.

Contribuições sociais/para a gestão
Como o artigo utiliza como cenário um contexto empresarial para analisar a possibilidade de se descobrir conhecimento dos dados IoT, utilizando uma estrutura de Data Lake, há como contribuição insumos teóricos para que os gestores obtenham novos insights sobre processos e produtos.

Palavras-chave: Internet das Coisas, Descoberta do Conhecimento, Data Lake

BUSINESS KNOWLEDGE DISCOVERY WITH IOT DEVICES GENERATED DATA: A DATA LAKE PERSPECTIVE

Study purpose
This study had as main goal to analyze, in the theoretical scope, the possibility of using a Data Lake structure, to execute a knowledge discovery process in data generated by IoT devices; using as background the business context for the analysis.

Relevance / originality
This article is relevant because no previous work was found that correlated the concepts or the correlated application of IoT devices, Data Lake and knowledge discovery, which makes the application and study of knowledge discovery difficult in this context.

Methodology / approach
A theoretical approach was used to revisit previous works that studied and discussed these concepts, isolated or in pair, and then to correlate the themes and analyze the feasibility of the knowledge discovery, using a Data Lake, with data from IoT devices.

Main results
Among the main results, IoT data is heterogeneous, produced at high speed and volume, the Data Lake structure allows optimizing data storage and retrieval in order to generate flexible inputs for knowledge discovery.

Theoretical / methodological contributions
This article provides as a theoretical contribution the joint study of the concepts Data Lake, IoT devices and knowledge discovery, analyzing the possibilities and difficulties, which will mainly give rise to practical studies on this joint approach.

Social / management contributions
Considering this study uses a business context as scenario to analyze the possibility of knowledge discovery with IoT data, using a Data Lake structure, there is as contribution the theoretical input for managers to obtain new insights about processes and products.

Keywords: Internet of Things, Knowledge Discovery, Data Lake

DESCOBERTA DE CONHECIMENTO EMPRESARIAL COM DADOS ORIUNDOS DE DISPOSITIVOS IOT: UMA PROPOSTA PARA O DATA LAKE

1 Introdução

Dispositivos *Internet of Things* (IoT), internet das coisas em tradução literal, são, conforme apresentado por Hussain *et al.* (2020), todos os dispositivos conectados a alguma rede de comunicação, por meio da qual enviam e recebem dados. Essa categoria de equipamentos foi impulsionada pelo avanço da tecnologia, que permitiu incluir módulos de processamento e sensores em diversos tipos de equipamentos, mas principalmente em razão da popularização das redes de internet através de acessos como o *Wi-fi*, 3G, 4G e 5G. Como apresentado por Mishra et al. (2014), em razão também da versatilidade desses equipamentos, o uso dos dispositivos IoT foi introduzido como ferramenta de coleta e comunicação nos mais diversos cenários e com diferentes objetivos.

Doan *et al.* (2020) apontam que apesar do grande número de dispositivos conectados à internet e emitindo dados, principalmente através de sensores embutidos, ainda existem desafios a serem superados para conseguir processá-los e utilizá-los de forma eficiente. Grueneber et al. (2019) elencam como o cerne do desafio para utilizar efetivamente esses dados ser o fato de que, além dos desafios no processo de coleta, processamento e utilização em treinamento de algoritmos, os dados oriundos dos dispositvos IoT só propiciam informações para a tomada de decisão, assim como a descoberta de conhecimento, após serem processados e analisados. Além disso, estes últimos autores ressaltam que, devido ao grande volume de dados, a análise só se torna viável com elemento inteligente de intermediação entre a coleta e a análise; elencam ainda que as técnicas de aprendizado de máquina são muito relevantes para se analisar os dados oriundos de dispositivos IoT.

Com a versatilidade dos dispositvos IoT, Cai *et al.* (2016) destacam que essa tecnologia tem sido muito utilizada no contexto empresarial, mas que, em razão de os dados serem coletados por sensores, há dificuldades em processá-los de forma que gerem valor às empresas. Indo além, os mesmo autores descrevem que o ambiente em nuvem oferece benefícios, mas também há desafios a serem enfrentados decorrentes da natureza dos dados oriundos desses dispositivos como: dados heterogêneos devido aos diferentes pontos de coleta, semântica de coleta, alto volume e presença de falhas nos dados. Jiang et al. (2014) corroboram que o dados IoT são heterogêneos, além do grande volume, e que é necessário haver uma ferramenta intermediária. Giebler *et al.* (2019) apontam a estrutura do *Data Lake* como uma alternativa para enfrentar os desafios impostos pela natureza dos dados gerados por dispositvos IoT, em razão do seu conceito ser baseado em uma maior flexibilidade para os dados gerenciados com o objetivo de permitir uma análise mais avançada.

Uma estrutura de *Data Lake* permite que quaisquer tipos de dados possam ser armazenados e extraídos para análise, o que é muito relevante no contexto empresarial, onde é necessário ser possível obter informações para gerar conhecimento sobre os componentes operacionais da empresa. Essa é uma estrutura que pode ser mais explorada no contexto do big data, principalmente para os dispositivos IoT, porque, como esses dispositivos geram dados diversos, em grande volume e velocidade, o agente intermediário deve estar preparado para centralizar e organizar dados com essas características. Inclusive, Mishra *et al* (2014) ressaltam que os dados IoT são gerados e transferidos de forma contínua, *stream*, sendo um

dos desafios para o gerenciamento e processamento do big data; principalmente para a descoberta de conhecimento.

De acordo com Kasemsap (2017), as empresas tem maior sucesso quando utilizam técnicas de visualização de dados e descoberta de conhecimento, pois conseguem extrair informações relevantes para gerar conhecimento no momento em que for necessário, a partir da análise de dados brutos — o que se correlaciona com os objetivos do *Data Lake*. Frawley e Piatetsky-Shapiro (1992), uns dos pesquisadores que foram pioneiros na abordagem da descoberta de conhecimento em bases de dados, explicam que os algoritmos para descobrir conhecimento identificam padrões relevantes e os apresentam de forma resumida e utilizável. Essas técnicas são valiosas ao serem implementadas visando a utilização em dados gerados por dispositivos IoT, mas, conforme apontado por Jiang *et al.* (2014), a rede desses dispositivos geram desafios para o processamento e análise mesmo em um ambiente em nuvem — evidenciando que não é qualquer estrutura que permitirá usufruir do potencial dos dados gerados por IoT.

Assim, este artigo tem como objetivo avaliar, no âmbito teórico, as possibilidades de utilizar uma estrutura de *Data Lake* para descobrir conhecimento sobre as atividades empresariais, a partir de dados gerados por dispositivos IoT. Este trabalho tem sua relevância no meio acadêmico, porque apesar de esses três temas terem sido amplamente abordados em trabalhos anteriores, poucos foram, os que correlacionaram as utilidades e possibilidades deles, principalmente visando a utilização empresarial; além de que cada um possuí, conforme já apresentado, seus desafios a serem enfrentados.

Este artigo é de natureza teórica e está organizado da seguinte forma: método de pesquisa bibliográfica utilizado para coleta do referencial teórico; extrato teórico, formado a partir dos trabalhos selecionados; discussão sobre os resultados da pesquisa e possibilidades de integração dos conceitos e estruturas (*Data Lake*, descoberta de conhecimento e dispositivos IoT); e conclusão.

2 Método

Considerando que o objetivo deste trabalho é avaliar as possibilidades de utilizar uma estrutura de *Data Lake* para descobrir conhecimento sobre as atividades empresariais, a partir de dados gerados por dispositivos IoT, primeiro realizou-se uma pesquisa bibliográfica sobre os trabalhos escritos anteriormente sobre o tema. Para isso, focando em termos que fossem relacionados às estruturas e conceitos abordadas neste artigo, foi utilizada a plataforma Google Acadêmico, realizando buscas com combinações das seguintes palavras-chave: *iot; data storage; knowledge discovery; problem; scheme; framework; mobile app; machine learning; Data Lake; cloud of things; data mining; e data visualization.*

Para cada das combinações de palavras-chave selecionou-se os artigos mais relevantes, em razão da estrutura do algoritmo *Page-rank* considerou-se como mais relevantes aqueles que apareceram nas duas primeiras páginas do sistema de busca; na Figura 1 abaixo encontra-se o quantitativo de artigos selecionados por combinação de palavras-chave. Com estes artigos selecionados, primeiro leu-se o título e resumo para avaliar a possibilidade de se enquadrarem no escopo do trabalho aqui apresentado; após isso, com uma ideia construída sobre cada um dos textos, filtrou-se aqueles que possuíam data de publicação maior do que dez anos anteriores, com exceção daqueles que foram pioneiros no assunto em razão de sua

posição histórica. Por fim, procedeu-se à leitura cuidadosa de cada uma das pesquisas selecionadas, realizando anotações sobre os trabalhos.

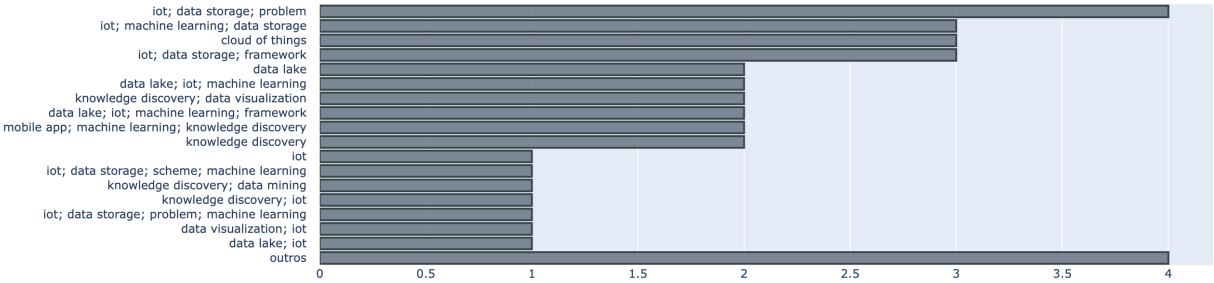

Figura 1. Quantidade de artigos selecionados por palavras-chave utilizadas na pesquisa

A fim de atingir o objetivo proposto nesta pesquisa, dividiu-se os trechos entre aqueles que seriam utilizados no Extrato Teórico (próxima seção), para contextualizar e conectar os temas ora em estudo, e os que seriam utilizados especificamente na parte de discussão, servindo de embasamento para as afirmações propostas. Isto porque, com essa pesquisa, pretende-se avaliar as possibilidades de utilizar uma estrutura de Data Lake, em um contexto empresarial, para gerar conhecimento a partir dos dados gerado por dispositivos IoT; assim, é imperioso estabelecer quais são os referenciais teóricos que estão servindo de base para as propostas de utilização. Na seção a seguir, será apresentado o referencial teórico selecionado para esta pesquisa.

3 Referencial Teórico

Mishra *et al.* (2014) definem os dispositivos IoT como qualquer objeto relacionado a algo existente no mundo real, que coleta dados e os transfere através de uma rede de comunicação. Parwekar (2011), ao apresentar IoT em seu trabalho sobre o início do evento chamado *Cloud of Things* que será tratado aqui posteriormente, vai além e define que os dispositivos IoT são aparelhos que coletam dados, por meio de sensores e atuadores embarcados na estrutura do objeto, e são conectados a uma rede com fio, ou sem, e podem utilizar os protocolos da internet para efetuar a comunicação. Esse último autor ressalta que o volume de dados gerado por esses dispositivos é abundante, o que pode ser um desafio, em empresas que tem a arquitetura preparada para realizar análise de dados estáticos, na transformação dos dados em valor.

Sobre o potencial da rede de comunicação entre dispositvos IoT, Peddoju e Upadhyay (2020), extendendo o que já foi conceituado, consideram que essa rede permite que dispositivos interajam entre si transferindo dados sobre eventos que permitem, em análise posterior ou até mesmo em tempo real, aprender sobre a entidade relacionado com o objeto. Indo além, estes autores consideram que esses aprendizados permitem que sejam tomadas ações preventivas ou otimizem atividades operacionais. Liang e Xiu (2018) ressaltam que é crescente o número de pessoas que utilizam dispositivos conectados à internet, principalmente em decorrência à disseminação de tecnologias como *Wi-Fi, 3G* e *4G*, ampliando a possibilidade de se realizar estudos com dados oriundos desses dispositivos e permite analisar características sobre o uso.

Conforme apresentado por Adi *et al.* (2020), uma das dificuldades para a utilização desses dados é que, com o grande número de dispositivos e, consequentemente, o volume de dados que são gerados, necessário é utilizar um sistema de processamento de dados inteligente e otimizado ao armazenamento dessas quantidade e frequência de dados. Para isso, esses autores consideram ser necessário que os dispositivos conectados e o aprendizado de máquina se complementem, aumentando a possibilidade de se extrair conhecimento dessa quantidade e variedade de dados. Doan *et al.* (2020), ressaltam que dentro desse contexto é imperioso, na montagem do *framework* de obtenção, armazenamento e recuperação dos dados se levar em consideração a multiplicidade de fontes de dados simultâneas.

A conexão de fontes de dados diversas, considerando uma rede de dispositivos IoT que coletam observações sobre diferentes entidades do mundo real, gera um conjunto de dados heterogêneos que, de acordo com Cai *et al.* (2016), é um dos desafios a serem enfrentados para conseguir realizar análises e obter informações. Segundo esses autores o processo de centralização, dos dados oriundos de diferentes dispositivos IoT, em uma forma que gere alguma utilidade é chamado de *Multi-source Data Fusion,* através de protocolos como BACnet que é focado na resolução dos dados para monitoramento de automações prediais. Este processo é essencial porque, segundo Majumdar e Mitra (2021), com o aumento da utilização e quantidade de dispositivos conectados surge o desafio de integrá-los e estruturá-los de forma que seja possível extrair valor dos dados coletados por dispositivos IoT; também ressaltam que, na atualidade, os dispositivos IoT são parte de um ecossistema, que compreende também análises em tempo real e aprendizado de máquina.

Kayes *et al.* (2020), em seus estudos sobre compressão de dados de séries temporal obtidos a partir de dispositivos IoT, para otimizar o espaço de armazenamento, ressaltam que ainda existem desafios para unir os dados, que na maioria das vezes são descentralizados e heterogêneos. Sobre o acesso aos dados IoT e suas inerentes dificuldades de armazenamento e coleta, Grueneber *et al.* (2019) corroboram que apesar da relevância e grande possibilidade de obter conhecimento através dos dados de IoT, a fim de gerar *insights* para a tomada de decisão, há desafios para obter os dados e aplicá-los em modelos de aprendizado de máquina. Esses autores abordam que, para conseguir utilizar esses modelos, é necessário que a análise seja feita em tempo real, ao invés de primeiro armazenar os dados e depois restaurá-los para análise; mas para isso, é necessário que o modelo seja treinado com dados rotulados. Pramukantoro *et al.* (2017) ressaltam que, devido à complexidade da estrutura de dados de dispositivos IoT, os bancos não relacionais são a escolha mais recomendada.

Peddoju e Upadhyay (2020), agora falando especificamente sobre a análise dos dados gerados por IoT, destacam que apenas com a realização dessa análise é que se pode extrair valor dos dados coletados por esses dispositivos, não apenas com visualização e monitoramento, a fim de influir no processo de decisão dentro da estrutura empresarial. Segundo esses autores, as vantagens da utilização de uma investigação mais aprofundada dos dados inclui a realização da análise para identificar entidades do mundo real, vinculadas aos dispositivos IoT componentes da rede de comunicação, que representem atividades essenciais no negócio e demonstrem potencial para melhoria. Mas, mesmo com essas considerações, os autores ressaltam que a utilização, nos dados de dispositivos IoT, de ferramentas como a visualização de dados, embora não sejam tão ótimas como a análise para identificar relevância e descobrir conhecimento, tem a sua importância para entender a rede de dispositivos, suas funções e objetivos.

Para extrair conhecimento de dados existe o processo conhecido como *knowledge discovery,* descoberta de conhecimento, o qual é descrito por Frawley e Piatetsky-Shapiro (1992) como a atividade de extrair informações que não eram conhecidas anteriormente, com alguma utilidade para o ambiente aplicado, como por exemplo o empresarial, e que não sejam triviais ao ponto de serem facilmente reconhecidas. Sobre o fato de a informação não ser trivial, pensando em um sistema de descoberta de conhecimento e que pode ser utilizado aqui como uma artifício explicativo, que ela é passível de ser descoberta quando sistema utilizado para ela tem alguma autonomia a fim de realizar cálculos e determinar se os achados são realmente conhecimento no contexto; além disso apontam que os dados brutos de bases de dados, assim como uma base de conhecimento e viés fornecidos pelo usuário, são os insumos desses sistema e o resultado de saída sistema é o conhecimento descoberto. Feyyad (1996), contemporâneo dos autores clássicos anteriores, compara a utilização da visualização de dados com o processo para a descoberta de conhecimento e pondera que há o fator humano como limitação do uso da visualização de dados neste processo, principalmente quando aplicado a um grande conjunto de dados.

Kasemsap (2017) contrapõe o autor anterior, no que tange a utilização de visualização de dados para a descoberta de conhecimento, afirmando que essas técnicas são uma forma fácil de entender como o sistema funcionado no todo; mas, em relação a conceituação da descoberta de conhecimento, corrobora que este processo tem como objetivo identificar conhecimentos que estão escondidos em grandes conjuntos de dados. Alhadari *et al.* (2014) ressaltam que o conhecimento pode ser obtido através de qualquer dados armazenado. Entretanto, Frawley e Piatetsky-Shapiro (1992), ressaltam que a descoberta de conhecimento em bancos de dados, que armazenem observações de entidades do mundo real, tem suas dificuldades porque normalmente eles são incompletos, com ruído e mudam constantemente — o que também apresentam como uma dificuldade para utilizar técnicas de aprendizado de máquina. Sobre a variedade de dados obtidas em um sistema IoT de *Big Data*, Mishra *et al.* (2014), relembram que dados podem ser coletados de forma estruturada, mas dados como os em formato log podem ter diversas peculiaridades, sendo uma dificuldade para a descoberta de conhecimento e gerenciamento dos dados.

Apresentando o termo *Big Data*, Gu *et al.* (2016) discorrem que com o aumento da utilização de sensores em aplicações há um aumento na capacidade de se coletar uma grande quantidade de dados, havendo também o desafio de obter conhecimento porque, como esses dados se enquadram no conceito de *Big* Data, eles são heterogêneos e o conhecimento está escondido dentro do grande volume de dados. Kiram *et al.* (2020), tratando sobre a análise do grande volume de dados gerados pelos dispositivos IoT, ressaltam que os sistemas que possuem esses dispositivos como forma de coleta são úteis e eficazes para identificar eventuais conhecimentos coletados; além de que a aplicação e resultado da análise sobre esses dados tende a resultar em melhoras para o contexto, por exemplo, um software ou processo. A análise do *Big Data* gerado pelos dispositivos IoT, segundo Mishra *et al.* (2014), tem como objetivo descobrir conhecimento, o que necessita de uma estratégia específica, e transformar esse conhecimento em formas de propiciar a tomada de decisão, principalmente relacionadas a sistemas em grande escala de automação em indústrias; para isso elencam que se deve modelar os dados, visualizá-los e analisar o que é apresentado nos dados identificando o conhecimento descoberto.

Sobre visualização de dados, uma das técnicas para analisar os dados IoT embora seja

controversa sua utilidade para a descoberta de conhecimento, Peddoju e Upadhyay (2020) apresentam que ela pode ser usada para fornecer auxílio aos tomadores de decisão no entendimento dos conceitos relacionados aos dados, mas em geral apenas de forma superficial; quando a visualização de dados é configurada de forma interativa, realizando novos processamentos de acordo com a necessidade do usuário, é possível ir além nas análises identificando padrões antes não identificados ou verificar formas de melhorar produtos e processos. De acordo com a definição de Kasemsap (2020) para o processo de tomada de decisão, a visualização de dados pode ser utilizada como uma ferramenta exploratória; mas, conforme ressaltado por Kasemsap (2017), há o desafio de se processar os dados gerados pelos dispositivos IoT, caracterizados como Big Data, para apresentá-los de forma visual, sendo necessário realizar uma etapa analítica anterior para reduzir a dimensionalidade dos dados. Wang e Zang (2020) apresentam que existem algoritmos baseados na computação em nuvem que facilitam o acesso e utilização dos dados armazenados nessa estrutura, como o *Hadoop Distributed File System* (HDFS), o que é utilizado no tratamento e armazenamento dos dados oriundos de dispositivos IoT.

Alhadari *et al.* (2020) ressaltam que a capacidade de computação interna dos dispositivos IoT não é suficiente para o processamento dos dados gerados, muito menos das computações necessárias para a análise do sistema a que estão inseridos como um todo, necessitando de um local externo para armazenar os dados, processá-los e realizar a análise. Para isso, elencam que a computação em nuvem é uma boa alternativa, permitindo extrair todo o potencial do *Big Data* e, utilizando a computação em nuvem, lidar com as representações do mundo real, captadas pelos sensores, e obter cenários realistas. Parwekar (2011), em seu artigo pioneiro na conceptualização do termo *Cloud of Things*, relembra que a Internet das Coisas começou de forma simples através de tecnologias como a *Radio Frequency Identification* (RFID), mas que para atingir todo o potencial dos dispositivos IoT, estruturados em uma rede de componentes, é necessário uma estrutura de armazenamento, que suporte tal fluxo e demanda de processamento, como a arquitetura de processamento baseada em nuvem. Mo (2019) corrobora com os autores anteriores no que tange a utilização da computação em nuvem aliada aos sistemas IoT, elencando que, devido ao aumento na utilização desses dispositivos e consequentemente a geração de um grande volume de dados, a maior parte dos dados é transferida para a nuvem, ambiente em que é possível obter armazenamento e processamento.

Aazam *et al.* (2014) discorrem ser a partir da integração com a arquitetura de armazenamento e processamento em nuvem que a utilização, de aplicações baseadas em dispositivos IoT, começa a ser enxergada como tendo uma possibilidade de integração horizontal, sendo que antes disso os dispositivos IoT eram tratados como silos verticais; servindo também como uma forma de tratar os dados de forma escalável e superar o desafio de utilização de dados heterogêneos. Em ambos os artigos, Aazam *et al.* (2014) e Alhadari *et al.* (2014), apontam que foi a partir da integração entre a Internet das Coisas e a computação em nuvem que surgiu o termo *Cloud of Things* (CoT). Alhadari *et al.* (2014) continuam que, com a integração dessas duas tecnologias, foi possível expandir as possibilidades das aplicações que utilizam dados de dispositivos IoT; isto porque, apesar destes dispositivos conseguirem obter dados que representem entidades do mundo real, isoladamente lhes faltava capacidade de armazenamento e processamento para dar valor aos dados. Parwekar (2011) ressalta ainda que a computação em nuvem retirou dos dispositivos IoT a necessidade de

suportar as tarefas de armazenamento e processamento, servindo apenas para coletar os dados, através de sensores e atuadores, e enviá-los à nuvem.

Jiang *et al.* (2014) discutem sobre o tipo de armazenamento que é necessário para suprir o volume de dados gerados por dispositivos IoT, ponderando que apesar de as tabelas relacionais ainda terem utilidade para várias aplicações, não tem capacidade para se adaptar à utilização como armazenamos de dados IoT. Estes autores descrevem que esses tipos de dados, como crescem de forma rápida, necessitam que a forma de armazenamento possa ser escalada horizontalmente, para isso existem as bases de dados NoSQL, as quais possuem funcionalidades que permitem escalar horizontalmente e indexação dos dados de forma mais eficiente, e ambientes distribuídos como o Hadoop. Mo (2019) descreve que o Hadoop é baseado principalmente no MapReduce e no sistema de armazenamento HDFS, o que faz com que seja uma escolha para quando é necessário uma forma dinâmica de armazenamento com computação distribuída, mas que seja possível entender como os dados estão distribuídos; além disso, indica que o Hadoop pode ser utilizado em diferentes tipos de aplicação para processar os dados e extrair informações. Outra estrutura para se utilizar com dados IoT, estrutura essa que, segundo Khine e Wang (2018) também pode ser aliada ao armazenamento NoSQL e ao Hadoop, é o *Data Lake* que, conforme apresentado por Oktan *et al.*, foi difundido com a ampliação da utilização de dispositivos IoT e permite que se extraia valor de dados em observações não antes exploradas devidamente.

A estrutura de *Data Lake*, segundo Grueneber *et al.* (2019), permite que sejam realizadas análises de forma flexível em conjuntos de dados complexos, como é o caso dos conjuntos de dados gerados por dispositivos IoT; segundo este autor, a primeira vez que o conceito de *Data Lake* foi utilizado para descrever uma estrutura que armazenasse os dados sem que houvesse algum tipo de tratamento, armazenando em sua forma bruta. Khine e Wang (2018) apresentam que o objetivo de um *Data Lake* é armazenar todos os dados gerados por uma empresa, independentemente de sua origem ou formato, em sua estrutura original deixando que os processamentos necessários para análise sejam realizados no momento da extração do banco de dados; isto permite que o conhecimento gerado tenha uma maior granularidade, além de, em combinação com a computação paralela, suportar as características dos dados identificados como Big Data. Giebler *et al.* (2021) apresentam que existem três classes de arquiteturas básicas de *Data Lake*: arquiteturas funcionais, focadas na extração e armazenamento dos dados; arquiteturas baseadas no refinamento dos dados, estruturando o *Data Lake* de acordo com o quão refinados são os dados brutos; e arquiteturas híbridas das classes anteriores. Estes últimos autores também elencam que a ingestão dos dados é flexível, podendo ser em fluxo contínuo, em lotes ou as duas formas combinadas.

Giebler *et al.* (2021) destacam também que a estrutura de *Data Lake* tem sido utilizada para superar as dificuldades das análises modernas com dados heterogêneos, mas que não existem muitos trabalhos acadêmicos que tratem sobre todo o fluxo de dados, da extração à análise, de forma compreensiva e que leve em consideração todas as questões relevantes para a aplicação. Importante é ressaltar que, conforme levantado por Cai *et al.* (2016), só a computação em nuvem não resolve problemas de processamento e armazenamento, incluindo escalabilidade e isolamento do armazenamento, gerados pela utilização de dispositivos IoT como ferramenta para a coleta de dados; isto principalmente devido à multiplicidade de origens dos dados, como em um sistema de diferentes sensores, que são heterogêneos e, além do grande volume, podem ser enviados em tempo real à nuvem.

4. Análise dos resultados e Discussões

Da pesquisa realizada e apresentada no extrato teórico desse artigo identificou-se, principalmente, os conceitos e objetivos dos dispositivos IoT, descoberta de conhecimento, computação em nuvem e *Data Lake*; além disso, coletou-se visões, algumas antagônicas, descritas na literatura acadêmica sobre as possibilidades e limitações dessas tecnologias e técnicas. A seção agora apresentada tem como objetivo resumir os conceitos apresentados, no extrato teórico, a fim de introduzir a discussão sobre a possibilidade e benefícios de utilizar uma estrutura de *Data Lake* para trabalhar com dados IoT.

Os dispositivos IoT geram grande volume de dados, em alta velocidade e, em uma rede de objetos dessa classe, uma grande variedade de dados devido à multiplicidade de fontes; sendo possível utilizar esses dados como insumo para a tomada de decisão em determinado contexto sobre alguma entidade do mundo real, em especial quando se utiliza objetos com sensores conectados em uma rede aprimorando as possibilidades de análise. Entretanto, principalmente em razão de cada dispositivo IoT gerar e enviar os dados com um formato diferente, há dificuldades na utilização desses dados devido à heterogeneidade, o que necessita de um tratamento anterior à análise um agrupamento, em um processo de fusão de dados, para começar a adentrar na busca por informações do sistema. Na literatura encontra-se a visão de que a análise é essencial para extrair valor dos dados, não sendo possível obtê-lo apenas com a visualização de dados.

A descoberta de conhecimento, *knowledge discovery*, é o processo de obter informações relevantes, não triviais, sobre os dados e que tenham alguma utilidade para o contexto, por exemplo a indústria de uma empresa; para este processo, há na literatura o entendimento de que o fator humano é um limitante para a utilização da visualização de dados como ferramenta. A realização da descoberta de conhecimento, com dados armazenados em bancos de dados, tem como dificuldade principal o fato de que os dados possuem ruído, o que pode levar a equívocos durante a análise, nos dados IoT adiciona-se como dificuldade o alvo volume e heterogeneidade. Também encontrou-se na literatura a visão de que, para os dados gerados pelos dispositivos IoT é necessário uma análise profunda a fim de identificar padrões e características das entidades relacionadas aos sensores, ou conjunto deles, em meio aos dados de diferentes fontes, possivelmente heterogêneos.

Na literatura também se encontrou que os dispositivos IoT possuem capacidades de armazenamento e processamento limitadas, sendo a computação em nuvem uma alternativa utilizada para se conseguir manipular e analisar o grande volume de dados dos dispositivos IoT , permitindo-se extrair todo o potencial do *Big Data* em uma união de ferramentas, IoT e computação em nuvem, que é denominada de *Cloud of Things* (CoT). Com a utilização dos ambientes em nuvem é possível realizar análises complexas com todo o conjunto de dados disponível, em vez de tratá-los como elementos isolados, silos, e analisá-los por partes; expande-se assim a utilização de aplicações baseadas em informações geradas por dispositivos IoT.

Uma das arquiteturas em nuvem encontradas na literatura é o *Data Lake*, o qual é apresentado como uma estrutura que armazene todos os dados, independentemente do formato, estruturado ou não, gerado por uma empresa; permitindo assim que se concentre todos os dados gerados no contexto empresarial de forma centralizada. Nesta estrutura, os

dados não são tratados antes do momento da análise e utilização, sendo armazenados da forma que são gerados; com isso há uma eficiência de recursos de processamento e permite que se realize análises mais complexas de um jeito mais simples, porque, além de haver um maior volume de dados a disposição, quando não armazenados em tabelas relacionais, a forma que se extrai os dados do *Data Lake* é estruturada no momento da utilização (*"schema on read"*).

A questão a ser aqui debatida é a possibilidade de utilizar, principalmente no contexto empresarial, uma estrutura de *Data Lake* como base para a descoberta de conhecimento com dados gerados por dispositivos IoT, tendo como base, para esta proposta, o estudo de artigos que abordaram o tema, selecionados conforme metodologia explicada anteriormente. Muito se vê na literatura acadêmica esses temas serem abordados, alguns, como a descoberta de conhecimento, desde o final do século passado; entretanto, poucos foram os acadêmicos que se debruçaram sobre a utilização de dados IoT para a descoberta de conhecimento, menos ainda foram os que abordaram o uso da estrutura de um *Data Lake* para armazenar e permitir análises sobre esses dados; inclusive, aqui nessa pesquisa não foram encontrados artigos que tratassem sobre o uso específico do uso do conceito de *Data Lake* para facilitar a descoberta de conhecimento. É provável que essa ausência de abordagens seja em decorrência de que cada um desses temas foi introduzido, cronologicamente, em momentos distintos, eles não são necessariamente temas contemporâneos.

É inegável a existência de benefícios em se utilizar a descoberta de conhecimento no contexto empresarial, uma vez que para a empresa toda informação que possa ser obtida e gere valor para a organização tem o potencial de favorecer o desempenho da atividade empresarial. Isto porque, a descoberta de conhecimento visa a obtenção de informações que não tivessem sido obtidas antes, não sejam triviais e tenham valor para a empresa; o que, em um contexto empresarial, pode significar insumos para as decisões que visem a melhoria de um processo ou, até mesmo, a indicação de algo que possa começar a ser feito como, por exemplo, o desenvolvimento de um produto que supra alguma necessidade dos usuários atuais com base em dados coletados. Neste contexto entra a possibilidade de se utilizar dispositivos IoT que possuem sensores para coletar dados do mundo real, incluindo o fluxo de pessoas em um ambiente, o funcionamento de máquinas em uma linha de produção, e o tempo decorrido por um veículo de entrega em seu percurso diário.

Na literatura se encontrou que os dispositivos IoT foram introduzidos inicialmente de forma simples, limitados pela tecnologia existente à época, mas que com o aprimoramento dessa tecnologia foi possível expandir os cenários de uso para esses dispositivos, aumentando as possibilidades de coleta de dados; com a popularização também das redes móveis, algo tão corriqueiro nos tempos atuais, os dispositivos começaram a se comunicar entre si, enviando dados coletados pelos sensores de um para que outro pudesse realizar determinada ação e, ainda, transferindo-os para um servidor com acesso remoto — surgindo assim os sistemas de dispositivos IoT integrados. Em um contexto empresarial talvez a aplicação mais simples de se pensar seja um complexo industrial no qual cada uma das máquinas possua sensores, coletando os dados sobre as entidades a eles relacionadas, adaptando os parâmetros de funcionamento de acordo com o coletado por outras máquinas na mesma linha de produção e, todas as máquinas, enviando simultaneamente esses dados para um servidor que permita a visualização e monitoramento dos equipamentos conectados. Mas, na realidade atual, com a integração com algoritmos de reconhecimento de imagens e inteligência artificial é possível

coletar dados sobre processos e fluxos que não necessariamente envolvam equipamentos, mas que mesmo assim tem o potencial de gerar conhecimento sobre as entidades e o contexto, por exemplo, as pessoas e como elas interagem com o ambiente permitindo que se analise possíveis pontos de melhoria.

Entretanto, conforme visto na literatura, com o grande volume de dados gerados por um dispositivo IoT, ainda mais por um sistema destes, a simples visualização e monitoramento dos dados não é suficiente para que se extraia valor dos dados, descobrindo conhecimento que possa gerar valor para a empresa e insumos para a tomada de decisão relacionada a processos e produtos. Como para isso é necessário que se realize análises complexas, inclusive, dependendo do contexto analisado, com a utilização de algoritmos de inteligência artificial, é necessário que se tenha uma estrutura robusta e preparada para coletar, armazenar e permitir uma extração eficiente dos dados, levanto em consideração o volume e variedade dos dados. Isto principalmente porque, em um sistema de dispositivos IoT, os dados são gerados por fontes diversas e precisam ser armazenados de uma forma que, independente de sua forma e origem seja possível extraí-los para uma análise conjunta dos dados, não analisando de forma isolada cada um dos dispositivos, afim de ampliar a busca por informações não triviais sobre as entidades e o sistema como um todo.

Seguindo essa lógica, de que os dados IoT precisam ser analisados em conjunto para se conseguir extrair todo o potencial, vê-se o *Data Lake* como útil para propiciar a descoberta de conhecimento com os dados IoT, uma vez que *Data Lake* é um conceito aplicado a uma estrutura, principalmente baseada na computação em nuvem devido à tendência de se utilizar esse ambiente para armazenar dados e aplicações. O *Data Lake* tem como princípio armazenar todos os dados gerados por diferentes fontes de um contexto, comumente o empresarial, em um só lugar de forma que estejam centralizados e possam ser extraídos quando necessários. Assim, como base nesse princípio, é viável considerar que um *Data Lake* possa ser aplicado para a descoberta de conhecimento com base em dados oriundos de dispositivos IoT, uma vez que, como é apresentado na literatura que ele serve para armazenar todos os dados gerados nos diferentes setores de uma empresa, pode-se fazer uma analogia de que, partindo do macro para o micro, o *Data Lake* armazenaria todos os dados gerados por sistemas de dispositivos IoT, cada setor da empresa seria um subsistema que representa um processo ou contexto e, dentro de cada sistema há os dispositivos com sensores embarcados para coletar dados das entidades do mundo real.

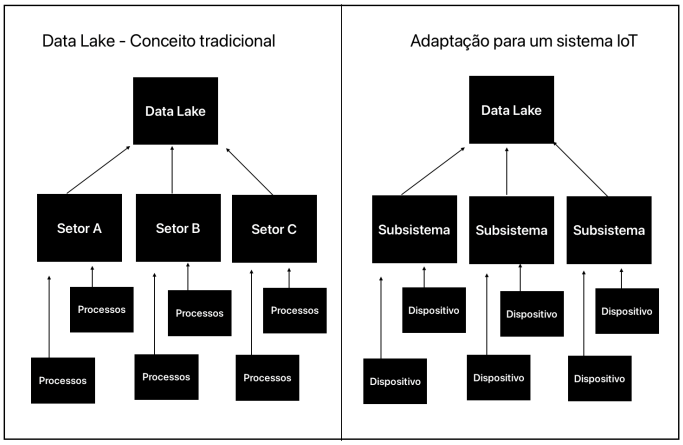

Figura 2. Adaptação do conceito tradicional de *Data Lake* para uso com sistemas IoT

Há de se fazer uma ressalva que o *Data Lake*, principalmente em um contexto empresarial, não comportaria única e exclusivamente os dados dos dispositivos IoT, mas também todos os dados que já são gerados e coletados na organização, dados estes que podem ser analisados em conjunto com aqueles, expandindo a análise e a possibilidade de se encontrar informações não triviais que não se levava em consideração anteriormente. Um exemplo, hipotético, dessa união seria coletar os dados dos parâmetros operacionais das máquinas utilizadas em uma linha de produção e agregá-los aos dados informados pelo setor de controle de qualidade, setor de vendas e o setor de atendimento ao cliente. Além dessa ressalva, há também que os dados de cada um dos sistemas de IoT e setores da empresa fossem armazenados em uma estrutura descentralizada; entretanto, neste estudo vê-se como uma das principais vantagens em se utilizar o *Data Lake* como base para a descoberta de conhecimento o fato de os dados estarem centralizados, o que é fundamental quando se quer analisar dados para extrair informações que não eram conhecidas antes — não se sabe, no momento de início da análise qual o resultado pretendido, o que vai de acordo com o disposto na literatura que um sistema de descoberta de conhecimento deve ser capaz de tomar decisões no meio do processo de acordo com as análises anteriores — entendendo a descoberta de conhecimento como um processo exploratório que pode necessitar de dados que não eram esperados ter relevância.

A outra vantagem vista para o *Data Lake* neste contexto é o fato de que o tratamento dos dados só é realizado no momento da extração para a análise, o que, além de resultar em uma economia computacional permite que se colete e armazene o maior número de dados. É claro que os dispositivos IoT geram um grande volume de dados, mas, caso fosse realizado o tratamento dos dados antes do armazenamento, correr-se-ia o risco de perder algum dado que seja relevante para uma descoberta de conhecimento futura, como um ruído, *outlier*, que poderia ser descartado, mas que significasse algum *lead* para uma análise maior sobre a entidade do mundo real ou, olhando para o macro, sobre o sistema. Por fim, entre a coleta de dados *stream* ou *batch*, vê-se como mais indicada a *stream*, mesmo para os dados que não são gerados por dispositivos IoT, como o de setores da empresa, porque, na realidade atual em que se tem à disposição algoritmos de inteligência artificial que podem tomar decisões, inclusive visando a descoberta de conhecimento, é ideal que os dados sejam os mais recentes para diminuir o tempo entre uma informação que seja relevante e gere a necessidade de recalibrar um algoritmo ou alterar um fluxo para otimizar os resultados do negócio.

5 Conclusões

Tendo como objetivo avaliar a possibilidade de utilizar uma estrutura de *Data Lake* para facilitar a descoberta de conhecimento com os dados oriundos de dispositivos IoT, este artigo se concentrou em revisar o que foi abordado na literatura sobre esses temas a fim de identificar as características, peculiaridades, possibilidades e dificuldades de cada um deles. Após isso, procedeu-se à apresentação do resultado dessa pesquisa, colocando de forma resumida os principais pontos do extrato teórico e, por última, a discussão sobre a possibilidade de integração desses três conceitos e tecnologias: descoberta de conhecimento; dispositivos IoT, *Data Lake*. Da pesquisa realizada, e consequente análise inferencial, conclui-se que o *Data Lake* é sim uma possibilidade para a descoberta de conhecimento com os dados IoT.

Isto porque, considerando que a descoberta de conhecimento é um processo exploratório, quanto maior a granularidade dos dados, maior a possibilidade de se ter como resultados informações relevantes e não triviais. E, considerando que os dispositivos IoT geram um alto volume de dados heterogêneos, de diferentes fontes, onde algo que possa ser previamente considerado com um ruído possa se revelar como um ponto de partida para a descoberta de um conhecimento relevante no cenário empresarial. Assim, como o *Data Lake* visa uma estrutura em que todos os dados gerados sejam armazenados da forma que são coletados, em um mesmo ambiente, há benefícios claros em utilizá-lo de forma a possibilitar e facilitar a descoberta de conhecimento com dados IoT; além de que os dados podem ser tratados, estruturados e selecionados apenas no momento da análise.

Não foi escopo desse artigo abordar a aplicação de qualquer um dos conceitos e tecnologias: descoberta de conhecimento, dispositivos IoT e *Data Lake*; mas analisar a possibilidade de integrá-los, de forma a gerar valor no contexto empresarial, e se o *Data Lake* é uma opção viável e benéfica para tal. Assim, para próximos trabalhos, pretende-se abordar a aplicação prática de um cenário em que se utilize o construído ao longo dessa pesquisa; além de pretender-se analisar os benefícios, nesse cenário, da utilização da computação em borda, *Edge Computing*, visando uma redução na latência de comunicação entre subsistemas IoT, assim como dispositivos, que necessitem comunicar dados para realizar ações ou afinar parâmetros, o que pode reduzir os ruídos nos dados gerados, em decorrência do atraso de comunicação, e ampliar as possibilidades de descoberta do conhecimento no contexto.

6 Referências

Aazam, M., Khan, I., Alsaffar, A. A., & Huh, E. N. (2014, January). Cloud of Things: Integrating Internet of Things and cloud computing and the issues involved. In *Proceedings of 2014 11th International Bhurban Conference on Applied Sciences & Technology (IBCAST) Islamabad, Pakistan, 14th-18th January, 2014* (pp. 414-419). IEEE.

Adi, E., Anwar, A., Baig, Z., & Zeadally, S. (2020). Machine learning and data analytics for the IoT. *Neural computing and applications*, *32*, 16205-16233.

Alhaidari, F., Rahman, A., & Zagrouba, R. (2020). Cloud of Things: architecture, applications and challenges. *Journal of Ambient Intelligence and Humanized Computing*, 1-19.

Cai, H., Xu, B., Jiang, L., & Vasilakos, A. V. (2016). IoT-based big data storage systems in cloud computing: perspectives and challenges. *IEEE Internet of Things Journal*, *4*(1), 75-87.

Doan, Q. T., Kayes, A. S. M., Rahayu, W., & Nguyen, K. (2020). Integration of iot streaming data with efficient indexing and storage optimization. *IEEE Access*, *8*, 47456-47467.

Feyyad, U. M. (1996). Data mining and knowledge discovery: Making sense out of data. *IEEE expert*, *11*(5), 20-25.

Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992). Knowledge discovery in databases: An overview. *AI magazine*, *13*(3), 57-57.

Giebler, C., Gröger, C., Hoos, E., Schwarz, H., & Mitschang, B. (2019). Leveraging the data lake: Current state and challenges. In Big Data Analytics and Knowledge Discovery: 21st International Conference, DaWaK 2019, Linz, Austria, August 26–29, 2019, Proceedings 21 (pp. 179-188). Springer International Publishing.

Giebler, C., Gröger, C., Hoos, E., Eichler, R., Schwarz, H., & Mitschang, B. (2021). The Data Lake Architecture Framework. *BTW 2021*.

Gu, Y., Jiang, H., Zhang, Y., Zhang, J. J., Gao, T., & Muljadi, E. (2016, September). Knowledge discovery for smart grid operation, control, and situation awareness—a big data visualization platform. In *2016 North American Power Symposium (NAPS)* (pp. 1-6). IEEE.

Grueneberg, K., Ko, B., Wood, D., Wang, X., Steuer, D., & Lim, Y. (2019, July). IoT Data Management System for Rapid Development of Machine Learning Models. In *2019 IEEE International Conference on Cognitive Computing (ICCC)* (pp. 59-63). IEEE.

Hussain, F., Hussain, R., Hassan, S. A., & Hossain, E. (2020). Machine learning in IoT security: Current solutions and future challenges. *IEEE Communications Surveys & Tutorials*, *22*(3), 1686-1721.

Jiang, L., Da Xu, L., Cai, H., Jiang, Z., Bu, F., & Xu, B. (2014). An IoT-oriented data storage framework in cloud computing platform. *IEEE transactions on industrial informatics*, *10*(2), 1443-1451.

Kasemsap, K. (2017). Knowledge discovery and data visualization: Theories and perspectives. *International Journal of Organizational and Collective Intelligence (IJOCI)*, *7*(3), 56-69.

Kasemsap, K. (2020). The fundamentals of neuroeconomics. In Foreign Direct Investments: Concepts, Methodologies, Tools, and Applications (pp. 99-130). IGI Global.

Kiran, S., Kumar, U. V., & Kumar, T. M. (2020, September). A review of machine learning algorithms on IoT applications. In *2020 International Conference on Smart Electronics and Communication (ICOSEC)* (pp. 330-334). IEEE.

Khine, P. P., & Wang, Z. S. (2018). Data lake: a new ideology in big data era. In *ITM web of conferences* (Vol. 17, p. 03025). EDP Sciences.

Liang, J., & Xiu, J. (2018, October). Prediction of Mobile APP Advertising Conversion Rate Based on Machine Learning. In *2018 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)* (pp. 284-2845). IEEE.

Majumdar, P., & Mitra, S. (2021). IoT and Machine Learning-Based Approaches for Real Time Environment Parameters Monitoring in Agriculture: An Empirical Review. *Agricultural Informatics: Automation Using the IoT and Machine Learning*, 89-115.

Mishra, N., Lin, C. C., & Chang, H. T. (2015). A cognitive adopted framework for IoT big-data management and knowledge discovery prospective. *International Journal of Distributed Sensor Networks*, *11*(10), 718390.

Mo, Y. (2019). A data security storage method for IoT under Hadoop cloud computing platform. *International Journal of Wireless Information Networks*, *26*(3), 152-157.

Parwekar, P. (2011, September). From internet of things towards cloud of things. In 2011 2nd international conference on computer and communication technology (ICCCT-2011) (pp. 329-333). IEEE.

Peddoju, S. K., & Upadhyay, H. (2020). Evaluation of IoT data visualization tools and techniques. *Data visualization: Trends and challenges toward multidisciplinary perception*, 115-139.

Pramukantoro, E. S., Yahya, W., Arganata, G., Bhawiyuga, A., & Basuki, A. (2017, October). Topic based IoT data storage framework for heterogeneous sensor data. In *2017 11th International Conference on Telecommunication Systems Services and Applications (TSSA)* (pp. 1-4). IEEE.

Wang, M., & Zhang, Q. (2020). Optimized data storage algorithm of IoT based on cloud computing in distributed system. *Computer Communications*, *157*, 124-131.