UNIVERSIDADE FEDERAL DE PERNAMBUCO

CENTRO DE ARTES E COMUNICAÇÃO

Tayná do Vale Tavares SIlva

DataOps como ferramenta de apoio ao gestor da informação

RECIFE

2024

UNIVERSIDADE FEDERAL DE PERNAMBUCO

CENTRO DE ARTES E COMUNICAÇÃO

GESTÃO DA INFORMAÇÃO

Tayná do Vale Tavares SIlva

DataOps como ferramenta de apoio ao gestor da informação

TCC apresentado ao Curso de Gestão da Informação da Universidade Federal de Pernambuco, Centro de Artes e Comunicação, como requisito para a obtenção do título de Bacharel em Gestão da Informação.

Orientador(a): Célio Andrade de Santana Júnior

RECIFE

2024

Ficha de identificação da obra elaborada pelo autor,
através do programa de geração automática do SIB/UFPE

Silva, Tayná do Vale Tavares.
DataOps como ferramenta de apoio ao gestor da informação / Tayná do Vale Tavares Silva. - Recife, 2024.
41

Orientador(a): Célio Andrade de Santana Júnior
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Pernambuco, Centro de Artes e Comunicação, Gestão da Informação - Bacharelado, 2024.

1. DataOps. 2. Gestão da informação. 3. Tecnologia . I. Santana Júnior, Célio Andrade de. (Orientação). II. Título.

020 CDD (22.ed.)

Serviço Público Federal
Universidade Federal de Pernambuco
Centro de Artes e Comunicação
Departamento de Ciência da Informação

FOLHA DE APROVAÇÃO

DATAOPS COMO FERRAMENTA DE APOIO AO GESTOR DA INFORMAÇÃO

TAYNA DO VALE TAVARES SILVA

Trabalho de Conclusão de Curso submetido à Banca Examinadora, apresentado no Curso de Gestão da Informação, do Departamento de Ciência da Informação, da Universidade Federal de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Gestão da Informação.

TCC aprovado em 17 de outubro de 2024

Banca Examinadora:

_____
CÉLIO ANDRADE DE SANTANA JÚNIOR - Orientador(a)
Universidade Federal de Pernambuco - DCI

_____
ANTÔNIO DE SOUZA SILVA JÚNIOR – Examinador(a) 1
Universidade Federal de Pernambuco - DCI

_____
JÔULDES MATOS DUARTE - Examinador(a) 2
PPGARQ/UFPE

RESUMO

O trabalho tem como objetivo apresentar o contexto de origem, o desenvolvimento, o uso, suas funcionalidades e um exemplo básico de código sobre o DataOps. Exemplo esse que demonstra como pode ser implementado dentro de uma organização. Além de trazer como o gestor da informação pode trabalhar com o sistema do DataOps e de que forma o mesmo está capacitado para gerir o uso. A metodologia adotada neste trabalho é uma combinação de uma revisão de literatura para a conceituação teórica do DataOps juntamente com uma apresentação de um experimento nos moldes de um *toy example* para apresentar um possível cenário de uso para o DataOps.

Palavras-chave: DataOps; Gestão da Informação; Tecnologia.

ABSTRACT

The work aims to present the context of origin, development, use, its functionalities and a basic code example about DataOps. An example that demonstrates how it can be implemented within an organization. In addition to bringing how the information manager can work with the DataOps system and how it is able to manage use. The methodology adopted in this work is a combination of a literature review for the theoretical conceptualization of DataOps together with a presentation of an experiment along the lines of a toy example to present a possible usage scenario for DataOps.

Keywords: DataOps; Information Management; Technology.

SUMÁRIO

1 – INTRODUÇÃO

O atual ambiente empresarial apresenta uma dinamicidade jamais observada em outros tempos. As empresas não só enfrentam os desafios inerentes aos próprios negócios, mas também, de fatores externos como novos produtos, novos perfis de clientes, ciclos de vida cada vez mais reduzidos de serviços, mudanças tecnológicas, um corpo de conhecimento crescente para soluções de problemas.

Villamil, Perez e Franco (2022) apontam que desde a virada do milênio as empresas se encontram naquilo que eles chamam de “mundo VUCA” onde a sigla significa:

V – Volatilidade (*Volatility*)

U – Incerteza (*Uncertain*)

C – Complexidade (*Complexity*)

A – Ambiguidade (*Ambiguity*)

O mundo VUCA é a representação da nova ordem de mercado que ocorre após o surgimento da Internet e dos mercados digitais. Essa metáfora sugere que as empresas estão imersas em um ambiente cada mais volátil, incerto, complexo e com múltiplas facetas de entendimento. Isso torna os negócios imersos na realidade VUCA mais vulneráveis a falhas. Para superar estes desafios são sugeridas uma série de mudanças organizacionais, dentre elas o maior aproveitamento das informações para responder a mudanças.

Mais recentemente, surgiu o conceito do mundo BANI, que segundo Martins (2021), é uma evolução do mundo VUCA e que veio a tona no período pós-pandêmico representa a nova configuração de evolução digital representada pela sigla onde:

B – Frágil (*Brittle*);

A – Ansioso (*Ansious*);

N – Não Linear (*Non-linear*);

I – Incompreensível (*Incomprehensible*).

Esses dois mundos de intensa transformação e efervescência para os negócios são materializados na quantidade de dados gerados por estas companhias, que

mesmo sendo de pequeno porte ainda se encontram imersas numa enxurrada de informação que se torna necessária para a tomada de decisão. Esta realidade é conhecida como aquilo que Ricciardo e colegas (2016) chamam de 7Vs do Big Data que são:

1. Volume (Quantidade de dados)
2. Variedade (Tipos de diferentes de dados)
3. Velocidade (Rapidez no processamento dos dados)
4. Veracidade (Segurança e confiança nos dados)
5. Valor (Utilidade dos dados)
6. Volatilidade (Tempo de vida útil dos dados)
7. Visualização (Capacidade de comunicação dos dados processados)

Os autores sugerem que a maioria das empresas passou a ter que lidar com um grande volume de dados dado pelo crescimento exponencial dos dados gerados e coletados por empresas modernas. Esta quantidade maciça de informação, proveniente de diversas fontes como mídias sociais, transações *online*, dispositivos de internet das coisas tem o seu processamento complexo e onde se faz um esforço cada vez maior para as empresas de análise eficiente. A capacidade de armazenar, gerenciar e processar esses grandes volumes de dados de forma eficaz é um desafio significativo.

Ainda segundo os autores, a variedade de dados que, além do volume, apresenta um desafio, uma vez que, tanto dados estruturados, quanto não estruturados e semiestruturados, precisam ser integrados e analisados conjuntamente. Dados de texto, vídeo, áudio, logs de máquinas e muitos outros formatos devem ser considerados, cada um exigindo diferentes técnicas de processamento e análise.

É apontado pelos autores que a velocidade de processamento e compreensão dos dados é um desafio. Em muitos casos, as empresas precisam de *insights* em tempo real ou quase real para tomar decisões rápidas e informadas. Isso exige uma infraestrutura de dados altamente responsiva e capaz de processar fluxos de dados em alta velocidade.

Outro ponto colocado se refere à veracidade dos dados, a qualidade e a precisão dos dados que são fundamentais. Dados imprecisos ou de baixa qualidade

podem levar a *insights* errôneos e decisões mal-informadas. As empresas enfrentam o desafio de validar, limpar e enriquecer os dados para garantir sua confiabilidade.

Outro V indicado pelos autores se refere ao valor dos dados, onde a extração dos dados de valor real e acionável é, talvez, o desafio mais crítico. Muitas empresas acumulam grandes quantidades de dados, mas não compreendem como identificar quais dados são úteis e como podem ser usados para impulsionar a inovação, otimizar operações ou melhorar a experiência do cliente.

Além disso, a complexidade tecnológica como fator determinante para o sucesso das empresas. A infraestrutura e as ferramentas necessárias para gerenciar eficientemente o ciclo de vida dos dados são complexas e em constante evolução. As empresas precisam de especialistas em dados, ferramentas avançadas e processos robustos para lidar com essa complexidade.

É apontado ainda as questões ligadas à governança e conformidade legal e normativas. Com novas regulamentações rigorosas de proteção de dados, como GDPR na Europa e LGPD no Brasil, as empresas enfrentam o desafio de garantir que seus processos de dados estejam em conformidade. Isso inclui a gestão de consentimentos, a proteção de dados pessoais e a garantia de que os dados são usados de maneira ética.

Nesse sentido se torna cada vez mais necessária uma área de atuação que ligue essas questões que envolvem os dados aos negócios, apoiado por uma estrutura tecnológica e é com esse conceito que surge a ideia do DataOps.

Ereth (2018) apresenta o DataOps, como uma abreviação do termo "*Data Operations*", que é uma noção oriunda das metodologias ágeis que visa melhorar a qualidade e reduzir o ciclo de análise de dados, integrando a engenharia de dados, a ciência de dados e a operação de dados em um fluxo de trabalho coeso. As empresas recorrem ao DataOps para resolver uma variedade de desafios relacionados à gestão de dados, à análise e à entrega de *insights* acionáveis.

Uma vez que a existência da má qualidade dos dados pode levar a tomada de decisões incorretas, o DataOps introduz processos para garantir a precisão, consistência e confiabilidade dos dados. O processamento de dados tradicional pode ser demorado em algumas situações e com a ajuda do DataOps passa a existir

um processo de automação e a entrega contínua acelerando o processamento de dados.

Munnapy e colegas (2020) apontam que a complexidade no gerenciamento de diversas fontes de dados devido à proliferação dessas fontes, torna difícil gerenciar e integrar estas informações, sendo assim um grande desafio que o DataOps resolve já que fornece ferramentas e práticas para lidar perfeitamente com diversas fontes de dados.

A conexão entre dados e informações no contexto de DataOps reside na transformação de dados brutos em intuições acionáveis, que são então utilizados para informar processos de tomada de decisão em vários domínios, desde mercados de trabalho até finanças e computação de ponta. Um problema frequente nas empresas é a existência de ruídos na comunicação entre os dados, ou seja, dificuldade na comunicação e acesso entre eles, o DataOps é focado em melhorar a comunicação entre os dados, assim, facilitando seu acesso e disponibilidade.

Os autores também apontam que as organizações geralmente têm dados armazenados em vários locais geograficamente separados, dificultando o acesso e a análise.

Dessa forma o DataOps é um sistema com um conjunto de práticas utilizadas com o objetivo de agilizar precisamente o processamento de dados, incluindo análises, controle de qualidade, acesso e integração. Um dos objetivos do DataOps é reduzir a quantidade de tempo da conclusão de um projeto de análise de dados, desde sua concepção de ideia até a construção de gráficos, modelos e tabelas. Seu fluxo de dados é constantemente gerenciado, o que em caso de anormalidade irá informar o time de analistas através de alertas, ou outras ações corretivas e preventivas.

O sistema proposto permite que a empresa tenha uma forma colaborativa de gerenciamento de informações otimizando a comunicação entre pessoas, integrando especialistas e automatizando o modo como os fluxos de dados ocorrem. Ferramentas baseadas em Big Data, machine learning e BI, são mais rápidas e precisas, além de contar com mais espaço para a transparência e a auditabilidade.

O DataOps atua por meio de fases, seu início se dá com com a coleta de dados, segue pelo armazenamento e vistoria de qualidade e desempenho, depois

ocorre a bateria de testes, avaliação preditiva e otimizações, em um ciclo de entregas contínuas.

Segundo a GAEA (2023):

> O primeiro procedimento é o de análise de dados, que passa pela estruturação até chegar à etapa de adaptação, uma das tarefas mais relevantes no processo. Tal prática compila, filtra e gerencia dados a fim de distribuí-los entre os próximos fluxos. Depois, são realizados testes até a entrega. Em seguida, a entrega retorna para uma segunda fase de adaptação, que faz o mesmo procedimento para identificar possíveis erros nos testes de pós-produção. Por fim, a administração é acionada para otimizar o uso dos recursos.

O DataOps surgiu com o desdobramento do Manifesto Ágil de 2001 que causou grandes mudanças na estrutura do desenvolvimento de softwares. O documento possui diretrizes para a criação de softwares, que serviram de orientação para o surgimento de plataformas de DataOps.

Devido a grande quantidade de dados presentes em uma empresa, a organização e análise desses dados para posteriormente torná-los em informações, se torna trabalhoso, demandando mais horas na tentativa de entender o que os dados podem representar ou indicar.

Existem vários problemas cujas resoluções são auxiliadas com o uso do DataOps garantindo que os processos ocorram sem problemas, sendo eles:

O problema da má qualidade dos dados pode levar a insights e decisões incorretas, onde o DataOps introduz processos para garantir a precisão, consistência e confiabilidade dos dados.

Na demora do processamento de dados tradicional o DataOps auxilia na automação e na entrega contínua acelerando o processamento.

Problemas nas condições que permitem incluir no processo de progresso do conhecimento científico podem ocorrer e a reprodutibilidade é essencial para reproduzir fluxos de trabalho de dados para fins de auditoria ou solução de problemas, o DataOps ressalta o controle de versão e a documentação para garantir a reprodutibilidade.

A complexidade no gerenciamento de diversas fontes de dados devido a proliferação de fontes de dados, torna difícil gerenciar e integrar os dados sendo assim um grande desafio, desta forma o DataOps fornece ferramentas e práticas para lidar perfeitamente com diversas fontes de dados.

A conexão entre dados e informações no contexto de DataOps reside na transformação de dados brutos em insights acionáveis, que são então utilizados para informar processos de tomada de decisão em vários domínios, desde mercados de trabalho até finanças e computação de ponta.

Todos os problemas citados trazem grande impacto nas organizações. Já que afetam diretamente os dados de uma organização e consequentemente afetam seu desempenho no geral. A partir dos dados se consegue uma apuração de operação sendo assim possível visualizar a atuação da organização, uma boa análise e um bom uso desses dados é essencial para o sucesso. Dito isso, torna-se necessário uma tomada de decisão sobre como um melhor manejo pode ser feito, gerando a necessidade e importância de se utilizar o DataOps.

Treleaven e col. (1982), dizem que a definição de *Data driven* pode ser vista como, “para denotar organizações de computação onde as instruções esperam passivamente que alguma combinação de seus argumentos fique disponível.”

Com a implementação do DataOps o tempo de insight é diminuído, não se perde tempo em problemas que ocorreram com o dado, a qualidade analítica melhora e se torna confiável e a eficiência do time se torna melhor com o aumento da agilidade, reuso e refatoração dos processos.

Segundo RIBEIRO e col. (2004), a refatoração é o modo de alteração da estrutura interna de um software tornando-o mais fácil e mais acessível a alterações, sem alterar o seu comportamento.

O Dataops é implementado no mundo atual em empresas que querem uma melhor e mais eficiente gestão da utilização e volume dos dados. Permitindo insights em tempo real com processos ágeis capazes de entregar insights em tempo real que amplia a capacidade de respostas e mudanças rápidas, aumentando também a democratização das informações, já que permite que os dados cheguem a todos os usuários e não se restrinja apenas aos cientistas de dados, com o aumento desse fluxo o foco estratégico pode ser ampliado facilitando a forma de lidar com as mudanças e oportunidades do mercado, aperfeiçoando a automação e comunicação dos fluxos de dados.

E considerando o Gestor da Informação neste contexto, a UFPE, no projeto pedagógico do curso de Bacharelado de Gestão da informação (GI), tem uma

perspectiva bastante próxima a essas necessidades acima explicitadas, quando afirma que:

> "Assim, enquanto ser humano, o gestor da informação é o profissional, com formação de caráter humanista, expressando a sua responsabilidade social e ética e perspectiva crítica frente à realidade social. Na esfera teórica e técnica seu objeto de gestão é a informação,e, assim, é responsável por coletar, selecionar, processar, armazenar, distribuir e avaliar o uso das informações, contribuindo, com seu trabalho, para o desenvolvimento socioeconômico, político e cultural da humanidade e, ainda, para a inclusão social dos menos favorecidos. A gestão da informação, portanto, diz respeito ao processo de gerir esse fluxo informacional (UFPE, 2020).”

1.1 - Competências do Gestor da informação

O gestor da informação tem competências e habilidades para atuar na área de tecnologia pois em sua formação é estudado várias disciplinas na área de tecnologia, como por exemplo base de dados, tópicos especiais em tecnologia, interação humano sistema, entre outras. Além da área de tecnologia, por ter um foco na gestão, o gestor da informação se torna um profissional perfeitamente competente para trabalhar com o DataOps já que o mesmo é focado em pessoas que trabalham com dados.

De acordo com a UFPE (2020) as competências tecnológicas que se esperam de um gestor da informação são: aspectos teórico-metodológicos e aplicados; conhecimento das Tecnologias para o incremento do uso eficiente da informação.

Ou seja, em sua formação o gestor da informação é capacitado para trabalhar no mundo das tecnologias focando principalmente no uso eficiente das informações.

As habilidades de um gestor da informação, segundo a UFPE(2020) são:

> Conhecer, refletir e aplicar teorias e modelos científicos de informação; Identificar, localizar e disponibilizar para seu cliente informações em diversos suportes; Identificar e explorar fontes de informação, o que requer habilidades em: navegação nas redes tradicionais e eletrônicas disponíveis, intercâmbio de informações entre sistemas de informação existentes, identificação de pessoas e organizações como fontes de informação, identificação, localização e análise de dados não cobertos por sistemas formais de informações; Projetar sistemas e repositórios de informação; Analisar, diagnosticar repositórios de informação, identificar problemas e projetar soluções; Avaliar a qualidade das fontes de informação, sob os seguintes parâmetros: exatidão, atualidade, abrangência, formatos

> disponíveis e orientada à necessidade do cliente. Adicionar valor ao processo de coleta de informações. Focar os parâmetros de qualidade do cliente. Antecipar as demandas de informação. Organizar e sistematizar a informação útil a cada cliente, utilizando-se dos processos de análise, interpretação e representação da informação.Coletar e conectar informações dispersas de modo a originar novas informações e conhecimentos. Utilizar a tecnologia como vetor para conectar pessoas, organizações, documentos e informações.

Dos problemas citados anteriormente, o gestor da informação tem plena capacidade de atuar no problema da má qualidade dos dados, já que tem a habilidade de gerir e identificar as necessidades dos dados assim garantindo uma maior a precisão e tornando esses dados mais confiáveis para serem utilizados dentro da organização.

A complexidade no gerenciamento de dados causada pelas diversas fontes de dados é um fator que o gestor também pode atuar, assim como na dificuldade existente na comunicação e acesso entre os dados, já que ambos podem ser resolvidos com a capacidade de gerir e organizar a informação.

Em síntese, um profissional formado em Gestão da informação, tem plena capacidade para trabalhar com o DataOps pois, além de ser um profissional com familiaridade com dados e informações, que possui entendimento sobre o seu funcionamento e necessidades, consegue juntamente com o DataOps, achar as melhores soluções e usabilidades dos dados da organização, havendo pouca dificuldade em manejar o sistema e utilizando da melhor forma toda sua capacidade. A formação de um gestor da informação e o conceito do DataOps entram em harmonia pelo fato que ambos possuem foco no dado.

Considerando todo este arcabouço, minha motivação enquanto gestora da informação para trabalhar com o DataOps é o que me fez seguir com este trabalho e enxergo como um tópico de relevância pois é algo que trabalha diretamente com o desempenho das organizações, onde o DataOps é uma das ferramentas que atuam nos dados brutos transformando-os em ideias e compreensões meneáveis, que serão utilizados para apontar os processos de tomada de decisão em vários domínios, assim facilitando o trabalho geral e influenciando em como a organização irá se portar no mercado ou em na forma que o funcionamento interno irá ocorrer.

A análise de dados é algo que faz parte do trabalho de um Gestor da informação, essa análise pode ser difícil pela complexidade dos dados ou pelo problema da má qualidade dos dados, como o auxílio do DataOps o trabalho da análise se torna mais eficaz, preciso e facilitado.

Nayak(1996) diz que: “A análise de dados é normalmente um processo iterativo e interativo envolvendo formulação de problemas, garantindo a qualidade dos dados, modelo construção, interpretação e pós-processamento dos resultados.”

Um profissional de gestão da informação pode colaborar com organizações utilizando DataOps de diversas formas já que o mesmo é uma metodologia que combina práticas de desenvolvimento ágil, DevOps e gestão de dados, assim melhorando a qualidade e velocidade da análise de dados. Com a implementação dessas práticas o profissional de gestão da informação ajuda com mais facilidade a organização a obter insights mais rápidos e precisos, aumentando a eficiência operacional, e respondendo de forma mais rápida e eficaz às mudanças nas demandas de negócios.

O objetivo geral do trabalho é apresentar os conceitos e finalidades do uso do DataOps no mundo organizacional em conjunto com a Gestão da informação mostrando algumas finalidades e tecnologias. Além de mostrar a importância atual do uso do DataOps e como ele pode trazer benefícios não só para o universo tecnológico, mas também para o Gestor da Informação que apresenta um perfil capacitado em atuar com o auxílio do DataOps.

Assim o trabalho possui os seguintes objetivos específicos:

- Apresentar o conceito de DataOps;
- Apresentar uma aplicação prática envolvendo o conceito de DataOps;
- Escrever um código como aplicação prática;
- Testar o código;
- Conceituar DataOps no dia a dia do gestor da informação

Então nesta perspectiva este trabalho também apresenta de forma prática o uso do DataOps e como o Gestor da informação pode utilizar e se beneficiar desse uso, assim ampliando sua capacidade de entendimento prático no trabalho em dados.

2 - Referencial Teórico

2.1 - Data Ops

Com a junção dos conceitos Agile, DevOps e dados se obtém a formalização do conceito de DataOps. A ligação entre o Manifesto Ágil e o DataOps pode ser vista na forma como o DataOps adota e adapta os princípios ágeis para o desenvolvimento e operações de dados.

O DataOps é influenciado pelos princípios e valores do Manifesto Ágil, onde os princípios são adaptados para o mundo dos dados, promovendo a agilidade, colaboração e entrega contínua de valor no contexto de desenvolvimento e operações de dados.

Segundo Tamburri et al. (2021):

> Assim como o DevOps, o DataOps visa combinar a produção, operação e entrega (de dados) em uma prática única e ágil que apoia diretamente funções de negócios específicas para melhorar qualidade, rapidez e colaboração e promover uma cultura de melhoria contínua. Uma metodologia DataOps combina e interconecta engenharia de dados, integração de dados, dados qualidade e segurança/privacidade de dados [8] para fornecer dados de suas fontes para a pessoa, sistema ou aplicativo que pode transformá-lo em valor comercial.

O foco do DataOps são pessoas que usam dados para analisar e criar modelos e não pessoas que estão acostumadas com a complexidade dos códigos, linguagens, ferramentas e etc. O DataOps promove que seja concebida uma plataforma de dados unificada e elimina silos, que são dados espalhados em vários locais, sem uma centralização.

Os *Pipelines* de dados, que são fluxos de processamento de dados automatizados, permitem a automatização das tarefas de transformação dos dados torna as pessoas livres para se concentrar em fazer as análises. Neste contexto, um dos problemas existentes é que os *Pipelines* de dados podem se tornar ineficientes e assim o DataOps introduz práticas para agilizar estes fluxos, tornando-os mais eficientes e menos sujeitos a erros.

Diferentemente do Manifesto Ágil citado anteriormente o “Manifesto DataOps” tem como foco que os indivíduos e as interações sendo mais importantes que os processos e ferramentas, a documentação e processos custosos são importantes

mas o analytic deve funcionar bem e se tornam mais relevantes, a colaboração com o cliente é mais relevante que a negociação de contratos, a experimentação, iteração e o feedback se tornam mais importantes que o design bem estruturado, ou seja o foco é o dado e sua qualidade. As operações devem ser multidisciplinares, que o conhecimento e o compartilhamento chegue em diversas áreas.

Quando se fala sobre desenvolver um sistema, as práticas DevOps e Agile são ideais para cumprir os requisitos, efetuar o desenvolvimento, testar e entregar um produto de maneira rápida e contínua, já quando se trata de sistemas analíticos, é necessário pensar sobre a arquitetura do fluxo de dados, já que os dados estarão continuamente entrando na alimentação de recursos utilizados pelo sistema alimentando o próprio sistema. É no fluxo de dados que se inicia a filosofia DataOps, É justamente aqui que se inicia a filosofia DataOps, que começa trazendo boas práticas que buscam diluir as barreiras e complicações que existem entre as as áreas de desenvolvimento e operações analíticas. (Lopes, 2019 apud Oliveira, 2020)

De acordo com o manifesto do DataOps, o mesmo possui 18 princípios sendo eles:

- Satisfaça continuamente seu cliente: Nossa maior prioridade é satisfazer o cliente por meio da entrega antecipada e contínua de insights analíticos valiosos, de alguns minutos a semanas.
- Análise de trabalho de valor: Acreditamos que a principal medida do desempenho da análise de dados é o grau em que análises criteriosas são fornecidas, incorporando dados precisos, sobre estruturas e sistemas robustos.
- Abrace a mudança: Acolhemos com satisfação as necessidades em evolução dos clientes e, na verdade, nós as abraçamos para gerar vantagem competitiva. Acreditamos que o método mais eficiente, eficaz e ágil de comunicação com os clientes é a conversa cara a cara.
- Uma equipe: As equipes analíticas sempre terão uma variedade de funções, habilidades, ferramentas favoritas e títulos. Uma diversidade de origens e opiniões aumenta a inovação e a produtividade.
- Interações diárias: Clientes, equipes analíticas e operações devem trabalhar juntos diariamente durante todo o projeto.
- Auto-organizar: Acreditamos que os melhores insights analíticos, algoritmos, arquiteturas, requisitos e designs emergem de equipes auto-organizadas.

- Reduzir o heroísmo: À medida que o ritmo e a amplitude da necessidade de insights analíticos aumentam cada vez mais, acreditamos que as equipes analíticas devem se esforçar para reduzir o heroísmo e criar equipes e processos de análise de dados sustentáveis e escaláveis.
- Refletir: As equipes analíticas devem ajustar seu desempenho operacional refletindo, em intervalos regulares, sobre o feedback fornecido por seus clientes, por eles próprios e pelas estatísticas operacionais.
- Analytics é o código: as equipes analíticas usam uma variedade de ferramentas individuais para acessar, integrar, modelar e visualizar dados. Fundamentalmente, cada uma dessas ferramentas gera código e configuração que descreve as ações tomadas com base nos dados para fornecer insights.
- Orquestrar: A orquestração completa de dados, ferramentas, código, ambientes e o trabalho das equipes analíticas é um fator-chave para o sucesso analítico.
- Torne-o reproduzível: resultados reproduzíveis são necessários e, portanto, versionamos tudo: dados, configurações de hardware e software de baixo nível e o código e configuração específicos para cada ferramenta no conjunto de ferramentas.
- Ambientes descartáveis: Acreditamos que é importante minimizar o custo para os membros da equipe analítica experimentarem, proporcionando-lhes ambientes técnicos fáceis de criar, isolados, seguros e descartáveis que reflitam seu ambiente de produção.
- Simplicidade: Acreditamos que a atenção contínua à excelência técnica e ao bom design aumenta a agilidade; da mesma forma, a simplicidade – a arte de maximizar a quantidade de trabalho não realizado – é essencial.
- Analytics é manufatura: *pipelines* analíticos são análogos às linhas de manufatura enxuta. Acreditamos que um conceito fundamental de DataOps é o foco no pensamento de processos que visa alcançar eficiências contínuas na produção de insights analíticos.
- A qualidade é fundamental: os *pipelines* analíticos devem ser construídos com uma base capaz de detecção automatizada de anormalidades (jidoka) e problemas de segurança em código, configuração e dados, e devem fornecer feedback contínuo aos operadores para evitar erros (poka yoke).

- Monitorar qualidade e desempenho: Nosso objetivo é ter medidas de desempenho, segurança e qualidade monitoradas continuamente para detectar variações inesperadas e gerar estatísticas operacionais.
- Reutilização: Acreditamos que um aspecto fundamental da eficiência da produção de insights analíticos é evitar a repetição do trabalho anterior do indivíduo ou da equipe.
- Melhorar os tempos de ciclo: Devemos nos esforçar para minimizar o tempo e o esforço para transformar uma necessidade do cliente em uma ideia analítica, criá-la em desenvolvimento, lançá-la como um processo de produção repetível e, finalmente, refatorar e reutilizar esse produto.

As organizações geralmente têm dados armazenados em vários locais, ruídos na comunicação entre os dados, ou seja, dificuldade na comunicação e acesso entre eles, a complexidade no gerenciamento de diversas fontes de dados devido a proliferação de fontes de dados, torna difícil gerenciar e integrar os dados sendo assim um grande desafio, problemas de reprodutibilidade, o processamento de dados tradicional pode ser demorado, o problema da má qualidade dos dados

2.2 Principais tecnologias adotadas para DataOps

O DataOps possui várias tecnologias e processos que facilitam o trabalho da equipe, sendo algumas tecnologias de gerenciamento de dados como, catálogos, virtualização, pipelines e gerenciamento de modelos de inteligência artificial. Existem também as tecnologias de controle e gerenciamento de versões, automação de testes, que utilizam Inteligência Artificial e Machine Learning para o auxílio dos processos e fluxos, evitando trabalho manual.

Os *pipelines* de dados permitem a automatização das tarefas de transformação de dados se concentrando em encontrar os melhores insights, um dos problemas existentes com isso é que os Pipelines de dados podem se tornar ineficientes e assim o DataOps introduz práticas para agilizar pipelines de dados, tornando-os mais eficientes e menos sujeitos a erros.

Existem 5 etapas que se podem dividir as ferramentas de apoio ao DataOps, sendo elas: O gerenciamento de controle de origem, automatização dos processos e fluxos de trabalho, adição de dados e testes lógicos, trabalho sem medo com

implantação consistente, implementação da comunicação e gerenciamento de processos.

- gerenciamento de controle de origem: um pipeline de dados nada mais é do que o código-fonte responsável por converter o conteúdo bruto em informações úteis. Podemos automatizar o pipeline de dados de ponta a ponta, produzindo um código-fonte que pode ser consumido de forma reproduzível. Uma ferramenta de controle de revisão (como o GitHub) ajuda a armazenar e gerenciar todas as alterações no código e na configuração para minimizar a implantação inconsistente.
- automatização dos processos e fluxos de trabalho: para que a metodologia DataOps seja bem-sucedida, a automação é a chave e requer um pipeline de dados projetado com flexibilidade de tempo de execução. Os principais requisitos para conseguir isso são serviços de curadoria de dados automatizados (como o Anders Pink), gerenciamento de metadados, governança de dados, gerenciamento de dados mestre e interação de autoatendimento.
- Adição de dados e testes lógicos: para ter certeza de que o pipeline de dados está funcionando corretamente, o teste de entradas, saídas e lógica de negócios deve ser aplicado (GitLab CI/CD). Em cada estágio, o pipeline de dados é testado quanto à 32 precisão ou desvio potencial junto com erros ou avisos antes de serem liberados para ter uma qualidade de dados consistente.
- Trabalho sem medo com implantação consistente: os profissionais de análise de dados temem a perspectiva de implantar mudanças que quebrem o pipeline de dados atual. Isso pode ser resolvido com dois fluxos de trabalho principais, que posteriormente se integram na produção. Primeiro, o pipeline de valor cria valor contínuo para as organizações(como SAS). Em segundo lugar, o pipeline de inovação assume a forma de novas análises em desenvolvimento que são posteriormente adicionadas ao pipeline de produção.
- Implementação da comunicação e gerenciamento de processos: notificações eficientes e automatizadas são essenciais em uma prática de DataOps. Quando alterações são feitas em qualquer código-fonte; ou quando um

pipeline de dados é acionado, falha, concluído ou implantado, as partes interessadas certas podem ser notificadas imediatamente. Ferramentas para permitir a comunicação entre as partes interessadas também fazem parte do conjunto de ferramentas (pense no Slack ou no Trello).

O conjunto de ferramentas ajuda na ascensão e reduz imprevistos na interoperabilidade. As tecnologias que funcionam no ambiente do DataOps são previstas como: local, nuvem, multi nuvem e híbrido. A Data Science Academy (2023), diz que algumas das principais ferramentas de DataOps incluem:

- Apache Airflow: um sistema de orquestração de pipelines de dados baseado em tarefas.
- AWS Glue: um serviço de ETL da Amazon que permite a criação, execução e gerenciamento de pipelines de dados.
- Talend: uma plataforma de integração de dados que oferece ferramentas para coletar, integrar e distribuir dados.
- Apache Nifi: um sistema de fluxo de dados de código aberto para automatizar a movimentação e o tratamento de dados.
- StreamSets: uma plataforma de gerenciamento de dados que permite a criação, execução e monitoramento de pipelines de dados.
- DataKitchen: uma plataforma de automação em DataOps.

O DataOps pode utilizar diversas linguagens de programação dependendo da necessidade e especificidades necessárias para o projeto, mas, existem algumas linguagens e tecnologias que são utilizadas com mais frequência, sendo elas:

- SQL (Structured Query Language): Que é utilizado para manipulação de banco de dados relacionais e é fundamental para operações de dados.
- Python: Uma das linguagens de programação mais populares na área de ciência de dados e análise de dados e é usada para desenvolver scripts, automação de processos, análise de dados e criação de pipelines de dados.
- R: Uma linguagem também conhecida na análise de dados e estatísticas, e em áreas acadêmicas e de pesquisa, para análise exploratória de dados e modelagem estatística.
- Shell Scripting: Linguagens de script como Bash são frequentemente usadas para automação de tarefas e para orquestração de processos em sistemas Unix/Linux.

- Scala: Uma linguagem usada em conjunto com o framework Apache Spark para processamento de grandes volumes de dados de forma distribuída.

Existem também, linguagens específicas que podem ser usadas em ferramentas e plataformas de DataOps, sendo elas linguagens de consulta para armazenamento de dados, linguagens de programação para desenvolvimento de plugins ou extensões em ferramentas de orquestração de pipelines de dados e etc.

As ferramentas de DataOps podem ser utilizadas para facilitar a sua implementação. “O Python é uma linguagem de programação amplamente usada em aplicações da Web, desenvolvimento de software, ciência de dados e machine learning (ML).” (AWS, 2024).

Segundo Borges(2014), as características da linguagem inclui a sintaxe clara e concisa que favorece a legibilidade do código fonte, assim o tornando mais produtivo. Ou seja, a linguagem Python é fácil de ser compreendida, facilitando o trabalho, além de ser bem estruturada permitindo que várias coisas sejam feitas com a linguagem.

O autor ainda afirma que, Python é um software de código aberto [...] o Python também é muito utilizado como linguagem script em vários softwares, permitindo automatizar tarefas e adicionar novas funcionalidades. A AWS, fala que, uma biblioteca é uma coleção de códigos usados com frequência que os desenvolvedores podem incluir em seus programas Python para evitar escrever o código do zero.

Em síntese as bibliotecas são uma forma de facilitar a criação do código, existem várias bibliotecas com várias funções que podem ser utilizadas para diversas aplicações, ciência de dados, machine learning entre outros.

2.3 Processos para DataOps

Segundo Oliveira(2020) as fases do processo do DataOps são na ordem: a análise, o desenvolvimento, a orquestração, o teste, a entrega, novamente a orquestração e por fim a administração.

Assim se percebe que de início ocorre a análise dos dados, passa pelo desenvolvimento até chegar ao processo de negócio onde os dados são organizados, se trata das exceções e as atribui ao próximo fluxo. Então o teste é feito para a entrega, após isso a segunda orquestração é feita permitindo que a organização monitore e controle os erros, por último a gestão otimiza o uso dos recursos.

“No DataOps o fluxo de valor e inovação funciona com "Pipeline de inovação" é o processo de introdução de novas ideias analíticas no pipeline de valor.” (Oliveira, 2020)

Sabe-se que o pipeline de inovação tem similaridade com o processo do DevOps mas no DataOps o processo é mais complexo já que possui duas orquestrações, na primeira o processo é responsável pela fábrica de dados onde o fluxo possui muitos passos e o responsável organiza os fluxos de dados e na segunda orquestração, além de tudo que ocorre na primeira ocorre também o monitoramento e controle de erros no processo de testes.

O DataOps pode ser utilizado em várias atividades, uma delas sendo na detecção de fraudes, especialmente no setor bancário, com as ferramentas do DataOps a abordagem tradicional da detecção de fraudes para cartão de crédito pode ser substituída e aumenta a velocidade da análise de dados. Com os insights avançados, os bancos evitam o uso de ligações e mensagens para seus clientes

O DataOps apresenta algumas utilidades específicas, como eficiência de produtividade, segurança e conformidade, transformação digital.

Segundo Pereira(2023) a eficiência em atividades essenciais com o DataOps é visto no:

- Alinhamento dos serviços;
- Automatização de tarefas;
- Avaliação produtiva;
- Construção de BI;
- Contribuição de diferentes setores;

- Potencialização da análise para deliberações;
- Redução de erros.

O DataOps pode ser utilizado com métodos para modelar dados, códigos e sistemas que integram dados, visando simplificar o armazenamento, utilizando um único ambiente virtual seguro para armazenar, já que várias fontes de dados atrapalham a produção. As tecnologias integradas e orientadas por dados, Tal qual as práticas padronizadas em conformidade, fazem com que os métodos ágeis de engenharia de software evoluam a cultura data driven corporativa.

Na gestão com o DataOps, a transformação digital possibilita a automatização dos acessos em fontes de dados na nuvem ou local. Dados privados e sensíveis são controlados com IA para permitir ou não acessibilidade conforme cada usuário, em uma eficaz governança. (PEREIRA, 2023).

2.4 Gestão da Informação no contexto a Big Data

Big Data é, segundo o Oracle(2023), “A definição de big data são dados que contêm maior variedade, chegando em volumes crescentes e com mais velocidade. Isso também é conhecido como os três Vs”.

A ACERT(2022) afirma que, “O big data não é apenas uma ferramenta de volume de dados. Ele é, na verdade, um mecanismo estratégico de análise. Isso porque, ao coletar, organizar e permitir a interpretação dos dados obtidos, é possível obter insights importantes sobre questões variadas.”

A Gestão da Informação (GI) trabalha com os processos do Big Data, de forma que a medida que as organizações lidam com volumes crescentes de dados, a necessidade de gerenciar, organizar e proteger esses dados se torna importante. A GI ajuda a definir quais, como e de onde os dados devem ser coletados, Isso inclui a identificação de dados relevantes garantindo que sua coleta seja feita de maneira ética de acordo com as regulamentações existentes.

O armazenamento e a organização na GI determinam como os dados são armazenados, organizados e catalogados. Envolvendo a seleção de sistemas de armazenamento, a criação de taxonomias e a implementação de soluções de gerenciamento de metadados.

Existem algumas áreas que mostram a relação e atuação da gestão da informação nos processos de Big data, sendo eles:

A Gestão da Informação estabelece processos que garantem a qualidade e integridade dos dados, incluindo a limpeza de dados, a resolução de inconsistências e a verificação de dados para garantir sua precisão e confiabilidade.

Medidas para proteger os dados são utilizados, sendo eles contra acessos não autorizados, perdas e violações, assim como também existe a garantia de que os dados sejam coletados, armazenados e processados.

A Gestão da Informação define como os dados são acessados, por quem é acessado e para qual propósito. Envolvendo a criação de interfaces de usuário, sistemas de busca e mecanismos de recuperação.

A análise e interpretação na Gestão da informação determina as ferramentas e técnicas que são mais adequadas para a análise dos dados. Estabelecendo padrões para a interpretação e apresentação dos resultados da análise.

A retenção e o descarte na Gestão da informação define políticas que determinam quanto tempo os dados devem ser retidos, quando e como devem ser descartados ou arquivados.

Na GI existem políticas e procedimentos que garantem os processos de Big Data estejam de acordo com as regulamentações internas e externas, incluindo a criação de estruturas de governança para supervisionar e orientar as atividades de Big Data.

Na colaboração e compartilhamento a GI facilita a colaboração entre equipes e departamentos definindo o compartilhamento dos dados na organização, externamente e internamente.

A Gestão da Informação atua em conjunto com os processos de Big Data, de forma que a garantia que os dados sejam coletados, armazenados, protegidos, analisados e utilizados de maneira eficaz e responsável, seja feita.

A GI e o Big Data trabalham com a coleta, armazenamento, processamento e análise de dados, mas têm focos e abordagens distintas.

Nas atividades principais da Gestão da informação, temos:

- Organização e catalogação de informações.
- Implementação de políticas de segurança e privacidade.
- Garantia da qualidade e integridade dos dados.
- Definição de políticas de retenção e descarte de dados.

Já nas atividades principais do Big Data temos:

- Coleta de dados de várias fontes, incluindo redes sociais, sensores, dispositivos, transações, etc.
- Armazenamento de grandes volumes de dados usando tecnologias como Hadoop e bancos de dados NoSQL.
- Processamento e análise de dados usando técnicas avançadas, como aprendizado de máquina e análise preditiva.
- Visualização de insights e resultados.

Assim pode-se perceber que a relação entre os dois é que a Gestão da Informação fornece a base para o gerenciamento eficaz de dados em uma organização, e o Big Data oferece as ferramentas e técnicas para analisar e extrair valor desses dados.

O gestor da informação por trabalhar com dados realiza todas as análises para obter as informações, pois é capacitado para interpretar os dados disponíveis. Silva, Moreira e Zawadzki (2019) diz que:

> A capacidade de gestão da informação (CGI) pode trazer o único insight necessário para a implementação de estratégias bem-sucedidas de Big Data. Definimos CGI como a capacidade da empresa de acessar dados e informações de ambientes internos e externos, mapear e distribuir dados para processamento, e permitir que se ajuste para atender as necessidades e direções do mercado.

Ou seja, com a gestão da informação as instituições passam a ter dados mais organizados e mais úteis para tomada de decisão.

Ribeiro (2014) aponta que “em função da interdisciplinaridade da nossa área, o cientista da informação é obrigado a lidar “com dados fragmentados de natureza empírica e teórica.” Além disso, o autor continua e complementa com a ideia de reformulação constante da Ciência”. Assim, o profissional da informação lida com os dados com base na visão da ciência da informação e métodos relacionados com o Big Data.

3 - Metodologia

A metodologia adotada neste trabalho é uma combinação de uma revisão de literatura para a conceituação teórica do DataOps seguida da apresentação de um experimento nos moldes de um *toy example* para apresentar um possível cenário de uso para o DataOps. A junção dessas etapas tinha como intuito alcançar o objetivo geral de apresentar os conceitos e finalidades do uso do DataOps no mundo organizacional, em conjunto com a Gestão da Informação, destacando suas finalidades e as principais tecnologias associadas.

A primeira etapa da pesquisa foi realizada por meio de uma revisão da literatura *ad hoc*, não sendo adotado nenhum procedimento sistemático, em fontes acadêmicas e técnicas sobre os temas de DataOps e Gestão da Informação. A pesquisa incluiu livros, teses, dissertações e publicações especializadas. Foram utilizadas bases de dados acadêmicas como Google Scholar, a fim de encontrar as fontes desejadas.

Segundo Michel (2000), a revisão de literatura é a abordagem mais utilizada em trabalhos monográficos. O objetivo deste método, é apresentar, explicar, discutir um tema com base em referências teóricas publicadas em veículos acadêmicos e técnicos. Segundo a autora, a revisão de literatura visa ampliar o conhecimento do autor sobre o assunto, sem a obrigatoriedade de aplicação dos seus resultados. a autora complementa, indicando que o método se mostra como uma importante ferramenta para a formação científica, que é o contexto deste trabalho de conclusão de curso.

Aqui, o foco foi analisar as definições, os conceitos centrais e as tecnologias envolvidas no assunto DataOps, bem como compreender a sua integração com a Gestão da Informação nas organizações. A pesquisa bibliográfica proporcionou uma base teórica, permitindo identificar as principais práticas, ferramentas e benefícios que a adoção do DataOps pode trazer para a organização no gerenciamento e uso eficaz da informação.

Durante essa fase, foram investigados estudos que abordam a implementação do DataOps e a sua contribuição para a otimização dos fluxos de dados, melhoria da qualidade da informação e integração entre diferentes equipes de TI e negócios. Além disso, foi explorado o papel da Gestão da Informação como facilitadora na

organização e processamento dos dados dentro do ciclo de vida proposto pelo DataOps. Essa revisão permitiu entender como essas abordagens contribuem para uma gestão mais ágil e eficiente da informação, alicerçando as bases teóricas do exemplo prático.

Na segunda etapa, foi realizado um *quasi-experimento* em forma de um *toy example*, que simularia uma situação de mundo real, em uma instituição financeira, o gestor da informação já identificou uma dificuldade no processamento de informação, devido a heterogeneidade de dados volumosos. No exemplo, foi concebida uma solução por um profissional com um perfil que conhece os dados que precisam ser processados (negócios), como esses dados precisam ser correlacionados (analista de dados) e como implementar uma infraestrutura tecnológica para realizar essa análise (especialista em tecnologia). A junção destas três competências formam o profissional apto a trabalhar com DataOps, entretanto, destaco que as duas primeiras (negócios e análise de dados) já são pontos fortes da formação de um gestor da informação. Assim, aqueles gestores que apresentam inclinação para a área da tecnologia, DataOps se mostra um caminho factível.

Neste contexto, essa pesquisa se define como uma pesquisa prática ou experimental, que segundo Michel (2000) ocorrem testes práticos de ideias e proposições discutidas na teoria. Segundo a autora, este tipo de pesquisa implica na simulação de ambientes reais, e tem como base a experimentação, na verificação de condições favoráveis à sua comprovação. A autora complementa que na área das Ciências Sociais, a pesquisa experimental pode ser aplicada em situações específicas para verificar, na prática, como se comportam as variáveis discutidas na teoria.

Neste pequeno exemplo, uma organização financeira que adotou práticas de DataOps integradas à sua Gestão da Informação. A escolha do experimento foi baseada na disponibilidade pública dos dados, na fácil compreensão do problema e no estreito contexto de negócio que simplificou a compreensão do problema, e não exigiu uma solução técnica demasiado elaborada. O código utilizado para o exemplo está presente nos apêndices A, B e C deste trabalho.

Por meio desta combinação da pesquisa bibliográfica e da pesquisa prática, a metodologia deste trabalho busca fornecer uma visão precisa e aplicada sobre o uso

do DataOps no contexto organizacional, destacando tanto os fundamentos teóricos quanto os resultados práticos da sua implementação.

4 - Implementação

Como já dito anteriormente, não existe uma única linguagem para ser usada no DataOps, a escolha depende do objetivo do projeto e as necessidades do mesmo, mas, no exemplo prático que será apresentado, foi utilizado o python por já haver uma maior afinidade com a linguagem. Com a linguagem python é possível desenvolver scripts, realizar automação de processos, análise de dados e criação de pipelines de dados.

Este é um exemplo inicial e básico de um pipeline de dados em Python. Dependendo das necessidades específicas do seu projeto é possível expandir e personalizar este pipeline com mais etapas, adicionando diferentes transformações de dados, manipulação de erros, logging, e outras funcionalidades que também serão apresentadas ao decorrer do capítulo.

4.1. Implementação de Pipeline de Dados em uma Empresa

A partir da análise de uma Empresa especializada em análise de dados para o setor financeiro, que enfrenta grandes dificuldades para automatizar e trazer qualidade no processamento de dados, o DataOps foi a melhor solução para a resolução dos seus problemas. O DataOps será utilizado para o gerenciamento de pipelines de dados que incluem, leitura, limpeza, transformação e exportação dos dados, tudo a partir de um código que funciona dentro do GitLab CI/CD que é uma ferramenta de CI/CD que são um conjunto de práticas que automatizam o processo de criação, teste e publicação de alterações de software. Integrada com o GitLab que suporta pipelines baseados em scripts Python. Os dados utilizados no processo são os dados de vendas dos seus clientes, das transações, produtos vendidos e preços e dados financeiros, receitas, despesas e lucros, que são os dados necessários para melhorar a eficiência operacional e a precisão dos dados, auxiliando na resposta às mudanças do mercado e na tomada de decisão da equipe financeira pois terão informações mais robustas e precisas.

A proposta encontrada envolve a implementação de um pipeline de dados em Python utilizando a biblioteca Pandas, com funcionalidades adicionais de logging para monitoramento e controle que é uma forma de DataOps, já que tem em seu foco a automação, eficiência e controle no gerenciamento de dados.

Todos esses códigos podem ser executados em qualquer ambiente Python com acesso às bibliotecas que serão utilizadas, em ambiente de computação em nuvem ou, se for necessário, um ambiente com funções mais complexas para realizar as automações e agendamentos de funções. Para o exemplo o indicado seria utilizar o GitLab CI/CD já que é integrado ao Git e permite uma automação e execução de pipelines diretamente do repositório.

O código presente no apêndice tem uma funcionalidade simples, ele cria uma pipeline de dados, automatiza o processamento de dados e implementa logs para o monitoramento das etapas. Uma pipeline serve para que dados de fontes diferentes sejam integrados e facilitem a análise, pois os dados serão exibidos em um novo documento de forma padronizada e simplificada. No código pode-se perceber uma transformação e um processamento sendo feitos, onde uma coluna foi convertida para datetime, que é um exemplo de processamento de fluxo para dados de pequeno porte, onde essa coluna irá ser utilizada para representar os eventos ou alterações que ocorreram em um certo período de tempo. Já o processamento feito, foi a criação de uma nova coluna que é utilizado para representar como os dados podem ser manipulados dentro da pipeline, como o código apenas apresenta um exemplo não fica claro o uso mas em um cenário real, essa criação de nova coluna para a manipulação de dados pode ser feita para cálculos, como margem de lucro onde ao invés da multiplicação ser por dois pode passar a ser por uma margem específica.

Existe uma função criada no código para definir a funcionalidade da pipeline, é possível ver que um documento vai entrar e outro vai sair, de forma prática, os documentos que irão entrar com dados serão limpos e processados, ou seja, a limpeza vai remover os dados faltantes no documento, o documento está em formato CSV que é “um arquivo de texto com formato específico para possibilitar o salvamento dos dados em um formato estruturado de tabela.” (GOOGLE, 2024).

O uso do Logging pode ser visto também e ele é utilizado para rastrear e monitorar as etapas do código, indicando os momentos que os dados foram carregados, processados e exportados. De forma mais visual, ocorre um registro quando os dados são carregados, processados e exportados, obtendo assim um rastro das etapas, que se por algum motivo não foi executada o log mostra até onde as etapas chegaram. No caso de haver alguma falha, esse rastro facilita a

localização do erro, já que além de demonstrar que ocorreu um erro, o logging indica a origem da falha, além de que esses avisos não quebram o fluxo do código, apenas mostra como está a sua execução. O código utilizado como exemplo não possui uso de alertas feitos em logging mas isso é algo que pode ser implementado, onde sempre que ocorre uma falha um alerta é disparado.

5 - Conclusão

Como profissional de Gestão da informação o trabalho apresentado é importante para entender uma das vertentes e ferramentas de uma das possíveis áreas de atuação do Gestor, onde se entende de uma forma abrangente as funcionalidades do DataOps e como pode vir a ser útil para trabalhos mais complexos com dados, disponibilizando um conhecimento não tão explorado até o momento.

Como o gestor da informação trabalha muito principalmente com dados, o conhecimento adquirido sobre o DataOps pode ser essencial para gerar um diferencial no seu âmbito profissional.

Em tese, o trabalho traz importantes informações e possíveis reflexões sobre a ferramenta e suas usabilidades.

Com mais tempo e conhecimento, a realização de trabalhos práticos e projetos na área são do meu interesse, pois estudando sobre o DataOps despertou interesse na área de manipulação de dados e na criação de projetos com a ferramenta, a automação, criação de pipelines de dados e monitoramento é algo que pode implementar muito na funcionalidade de uma organização de forma que o trabalho de todos pode ser facilitado.

Todos os profissionais de Gestão da informação que querem dar ênfase em tecnologia devem procurar cursos para se aprofundar e melhorar suas competências. O curso de graduação apresenta ferramentas que podem ajudar no início e dar uma base de por onde se deve começar, mas para estar apto ao mercado de trabalho e entender de verdade as tecnologias e a abrangência do DataOps, é necessário um estudo aprofundado que não está disponível na graduação. A prática é o mais importante para o aperfeiçoamento, como DataOps ainda está crescendo, existem poucas coisas sobre, sendo necessário estudar muito por fontes que estão em outras línguas, principalmente inglês. Então, além do conhecimento técnico em tecnologias e linguagens de programação, se faz necessário que se tenha conhecimento na língua inglesa para que os estudos sejam facilitados.

Além de estudo e prática em projetos, é importante que exista contato com profissionais da área para que se entenda melhor o funcionamento dentro de

empresas, assim obtendo conhecimento de pessoas que já possuem experiência e podem passar esse conhecimento de forma que seja mais simples e menos assustador todo o processo do mercado de trabalho e o trabalho dentro das organizações.

REFERÊNCIAS

ACERT. Big Data: o que é, para que serve, como aplicar e exemplos. Disponível em: https://acertbr.com.br/big-data-o-que-e-para-que-serve-como-aplicar-e-exemplos/. Acesso em: 25 out. 2023.

AWS. O quê é Python. (2024) Disponível em: https://aws.amazon.com/pt/what-is/python/ Acesso em: 25 out. 2023.

BARBOSA, Wellington Luiz; LYRA, Roberto Shayer. Governança de Dados. Brasília: Enap, 2019. 49 p. Disponível em: https://repositorio.enap.gov.br/bitstream/1/5008/2/M%C3%B3dulo%202%20-%20Princ%C3%ADpios%2C%20import%C3%A2ncia%20e%20desafios%20do%20Gerenciamento%20de%20Dados.pdf. Acesso em: 25 out. 2023.

BORGES, Luiz Eduardo. Python para desenvolvedores. São Paulo: Novatec, 2014. Disponível em: https://novatec.com.br/livros/python-para-desenvolvedores/. Acesso em: 14 out. 2023.

DATAOPS Manifesto. Disponível em: https://dataopsmanifesto.org/en/. Acesso em: 20 out. 2023.

DATA SCIENCE ACADEMY. O Que é DataOps? Um Exemplo de Caso de Uso. Disponível em: https://blog.dsacademy.com.br/o-que-e_dataops/. Acesso em: 27 out. 2023.

ERETH, Julian. DataOps: towards a definition. In: Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC 2018), v. 2191, p. 104-112, 2018. Disponível em: https://www.aclweb.org/anthology/L18-1036. Acesso em: 20 nov. 2023.

GAEA. DataOps: o que é, como funciona e como implementá-lo?. o que é, como funciona e como implementá-lo?. Disponível em: https://gaea.com.br/dataops/. Acesso em: 27 nov. 2023.

GOOGLE. Como fazer upload de dados offline para o Google Ads usando arquivos CSV. Disponível em: https://support.google.com/google-ads/answer/9004364?hl=pt-BR. Acesso em: 10 set. 2024.

GONÇALVES, Marcio; PEREIRA, Cláudio Lopes. Gestão da informação na era do Big Data. Revista Dissertar, [S.L.], v. 1, n. 2425, p. 71-80, 1 jun. 2016. Associação de Docentes da Estácio de Sá (ADESA). Disponível em: http://dx.doi.org/10.24119/16760867ed11239, Acesso em: 11 out. 2023.

GUPTA, Ritesh. Components of the DataOps toolchain and best practices to make it successful. 2019. Disponível em: https://www.ibm.com/blog/components-of-the-dataops-toolchain-and-best-practices-to-make-it-successful/. Acesso em: 20 nov. 2023.

MICHEL, Maria Helena. Metodologia E Pesquisa Científica Em Ciências Sociais . Editora Atlas SA, 2000.

MARTINS, Dora. *Do Mundo VUCA ao Mundo BANI: impactos na gestão de empresas e na gestão de pessoas*. 2021. Disponível em: https://www.universidade.com. Acesso em: 14 nov. 2023.

MUNAPPY, Aiswarya Raj et al. From ad-hoc data analytics to DataOps. In: Proceedings of the International Conference on Software and System Processes (ICSSP 2020). 2020. p. 165-174. Disponível em: https://ieeexplore.ieee.org/document/9130279. Acesso em: 14 nov. 2023.

NAYAK, Richi. Intelligent Data Analysis: issues and challenges. 2002. 12 f. Tese (Doutorado) - Curso de Information Systems, Queensland University Of Technology, Austrália, 2002.

OLIVEIRA, Haila Bianca dos Santos de. DATAOPS: o novo paradigma de metodologia ágil. 2020. 39 f. TCC (Doutorado) - Curso de Tecnologia em Redes de Computadores, Instituto Federal de Educação, Ciência e Tecnologia do Amapá, Macapá, 2020. Disponível em: http://repositorio.ifap.edu.br/jspui/bitstream/prefix/271/1/OLIVEIRA%20%282020%29%20DataOps%20O%20Novo%20Paradigma.pdf. Acesso em: 25 out. 2023.

ORACLE. O que é Big Data? Disponível em: https://www.oracle.com/br/big-data/what-is-big-data/. Acesso em: 30 nov. 2023.

PANDAS. User Guide. Disponível em: https://pandas.pydata.org/docs/user_guide/index.html#user-guide. Acesso em: 15 fev. 2024.

PEREIRA, Flávia. DataOps: entenda como implementar na sua empresa. Entenda como implementar na sua empresa. Disponível em: https://www.dataex.com.br/dataops-entenda-como-implementar-na-sua-empresa/. Acesso em: 30 nov. 2023.

RIALTI, RICCARDO et al. Big data oriented business models: the 7vs of value creation. In: XXVIII Sinergie Annual Conference: Management in a Digital World. Decisions, Production, Communication. http://dx. doi. org/10.7433/SRECP. EA. 2016. Acesso em: 20 nov. 2023.

RIBEIRO, C. J. S. Big Data: os novos desafios para o profissional da informação. Informação & Tecnologia, [S. l.], v. 1, n. 1, p. 96–105, 2014. Disponível em: https://periodicos.ufpb.br/index.php/itec/article/view/19380. Acesso em: 11 mar. 2024.

RIBEIRO, Márcio André Brandão; LESSA, Andé Filipe Lourenço; SOUSA, Bruno Alexandre Silva de; EIRAS, João Carlos Martins; MACHADO, Nuno Miguel das Neves; MORERIRA, Rodrigo Manuel Lopes de Matos. Engenharia de software: refactoring. Porto: Eup, 2004. Disponível em: https://paginas.fe.up.pt/~ei02017/docs/relatorio_es.pdf. Acesso em: 20 nov. 2023.

SILVA, Igor Lopes da; MOREIRA, André Luiz; ZAWADZKI, Moisés. Capacidade de Gestão da Informação e Implementação de Estratégias de Big Data. *Revista de Administração de Empresas*, v. 61, n. 5, p. 1-12, 2021. Disponível em: https://www.scielo.br/j/rae/a/WK3bSK9mMXmPdB8tdZVjLTB/?lang=pt. Acesso em: 30 nov. 2023.

TAMBURRI, Damian A. et al. DataOps for Societal Intelligence: a data pipeline for labor market skills extraction and matching. Países Baixos: Jads, 2020. 394 p. Disponível em: https://arxiv.org/pdf/2104.01966.pdf. Acesso em: 30 out. 2023.

TRELEAVEN, Philip C.; BROWNBRIDGE, David R.; HOPKINS, Richard P. Data-Driven and Demand-Driven Computer Architecture. ACM Computing Surveys, New York, v. 14, n. 1, p. 93–143, mar. 1982. Disponível em: https://doi.org/10.1145/356869.356873. Acesso em: 27 nov. 2023.

Universidade Federal de Pernambuco. Projeto Pedagógico do Curso de graduação em Gestão da Informação. 2020. Disponível em: https://www.ufpe.br/documents/39179/0/Perfil_103.2.pdf/a5e74b1b-c00e-4b15-8b66-bae8610efb55. Acesso em: 23 out. 2023.

VILAMIL, Eduardo; PEREZ, Alvaro; FRANCO, José. O mundo VUCA: uma abordagem para as empresas do século XXI. *Semillas del Saber*, 2022. Disponível em: https://www.semillasdelsaber.com.br/o-mundo-vuca. Acesso em: 20 nov. 2023.

VINAY SAJIP. HowTo - Logging. Oregon: Python, 2023. Disponível em: https://docs.python.org/pt-br/3/howto/logging.html. Acesso em: 17 fev. 2024.

APÊNDICE A: O CÓDIGO DESENVOLVE E IMPLEMENTA UM PIPELINE DE DADOS COM PROCESSAMENTO AUTOMATIZADO, CARREGA DADOS DE ARQUIVOS CSV, LIMPA, PROCESSA, TRANSFORMA E SALVA OS DADOS PROCESSADOS EM UM NOVO ARQUIVO, UTILIZANDO LOGS PARA MONITORAR AS ETAPAS DE EXECUÇÃO.

```
import pandas as pd
import logging

# Configuração do logger
logging.basicConfig(level=logging.INFO)

# Primeira fase: Carregar os dados
def load_data(file_path):
    logging.info("Carregando os dados do arquivo: %s", file_path)
    return pd.read_csv(file_path)

# Segunda fase: Limpar os dados
def clean_data(data):
    logging.info("Limpando os dados...")
    # Exemplo de limpeza simples: remover linhas com valores ausentes
    cleaned_data = data.dropna()
    return cleaned_data

# Terceira fase: Processar e transformar os dados
def process_transform_data(data):
    logging.info("Processando e transformando os dados...")
    # conversão de uma coluna para formato datetime
    data['data'] = pd.to_datetime(data['data'])
    # adiciona nova coluna
    data['nova_coluna'] = data['coluna_existente'] * 2
    return data

# Quarta fase: Salvar os dados processados
def save_data(data, file_path):
    logging.info("Salvando os dados processados em: %s", file_path)
    data.to_csv(file_path, index=False)
    logging.info("Exportação concluída.")

# Função principal que executa o pipeline de dados
```

```
def main():
    # Caminho do arquivo de entrada
    input_file = 'dados.csv'
    # Caminho do arquivo de saída
    output_file = 'dados_processados.csv'

    # Fase 1: Carregar os dados
    data = load_data(input_file)

    # Fase 2: Limpar os dados
    cleaned_data = clean_data(data)

    # Fase 3: Processar e transformar os dados
    processed_data = process_transform_data(cleaned_data)

    # Fase 4: Salvar os dados processados
    save_data(processed_data, output_file)

if __name__ == "__main__":
    main()
```