1. Introdução

A evolução e popularização dos recursos computacionais e da digitalização tornaram essencial o desenvolvimento de softwares capazes de lidar com o crescente volume e complexidade das informações empresariais. A integração eficiente de dados passou a ser um fator estratégico, agregando valor a diversas áreas e tornando-se fundamental para que empresas sejam competitivas no mercado.

Com o avanço da tecnologia, as organizações adotam múltiplas ferramentas de diferentes fornecedores, cada uma com sua estrutura e modelo de dados próprios. Essas soluções, embora otimizem processos internos, criam um novo desafio: a fragmentação da informação. Para que os dados gerados por diferentes sistemas possam ser aproveitados de forma unificada, é necessário estabelecer mecanismos de integração que permitam sua padronização e relacionamento inteligente.

Como apontado por Ziegler e Dittrich (2004), a necessidade de consolidar dados provenientes de diversas aplicações tornou-se um desafio recorrente no contexto empresarial. Estruturas distintas, granularidades variadas e formatos heterogêneos dificultam a obtenção de uma visão única e coesa das informações. Weidman e Arrison (2010) reforçam essa ideia, destacando que a combinação desses dados deve permitir consultas complexas e análises avançadas, sem comprometer a qualidade e a confiabilidade da informação.

Neste contexto, o uso de bancos de dados em grafos aliado a técnicas de aprendizado de máquina (machine learning) surge como uma solução inovadora. Modelos baseados em grafos permitem representar de forma dinâmica e flexível as relações entre diferentes fontes de dados, enquanto algoritmos de aprendizado de máquina possibilitam a descoberta automática de padrões e conexões não explícitas. Dessa forma, é possível criar um sistema de integração que não apenas unifica as informações, mas também gera novos insights a partir delas.

1.1 Descrição da empresa

1.1.1 Identificação

1.1.2 Segmento de atuação

1.1.3 Descrição do problema

A fragmentação da informação nas empresas é um problema crítico que impede a obtenção de uma visão unificada e coesa dos dados. Com a adoção de múltiplas ferramentas de diferentes fornecedores, cada uma com sua própria estrutura e modelo de dados, as organizações enfrentam dificuldades para consolidar e integrar essas informações. Estruturas distintas, granularidades variadas e formatos heterogêneos tornam complexa a tarefa de padronizar e relacionar os dados de maneira inteligente. Esse cenário resulta em desafios significativos para a realização de consultas complexas e análises avançadas, comprometendo a qualidade e a confiabilidade das informações.

A crescente digitalização das operações empresariais tem levado à geração de grandes volumes de dados provenientes de diversas fontes, incluindo sistemas internos, planilhas, arquivos e bancos de dados de diferentes fornecedores. A necessidade de integrar essas informações de forma eficiente tornou-se um desafio essencial para empresas que buscam otimizar seus processos, melhorar a tomada de decisão e manter-se competitivas no mercado.

Bancos de dados relacionais tradicionais, apesar de amplamente utilizados, apresentam limitações quando se trata da modelagem e análise de dados interconectados. A necessidade de realizar múltiplas consultas para recuperar informações distribuídas em diferentes tabelas pode resultar em consultas lentas e ineficientes, especialmente quando os dados crescem em volume e complexidade. Além disso, a rigidez do esquema relacional dificulta a adaptação a mudanças nos formatos e nas relações entre os dados, o que é um problema relevante quando lidamos com fontes diversas e dinâmicas.

Por outro lado, bancos de dados em grafo surgem como uma alternativa viável para esse cenário. Diferentemente dos bancos relacionais, os grafos representam dados como nós (entidades), arestas (relacionamentos) e propriedades, permitindo uma modelagem mais flexível e intuitiva de dados interconectados. Esse modelo elimina a necessidade de consultas complexas, tornando as consultas mais rápidas e eficientes.

A aplicação de técnicas de Machine Learning, como regressão linear e classificação, pode agregar ainda mais valor ao modelo de grafos. Algoritmos de aprendizado de máquina podem ser utilizados para identificar padrões, prever tendências e gerar novas conexões entre os dados, possibilitando análises mais profundas e estratégicas. Dessa forma, a combinação de bancos de dados em grafo e Machine Learning permite não apenas consolidar e integrar informações de diversas fontes, mas também descobrir relações que não foram previamente definidas, trazendo novas percepções e melhorando a capacidade de tomada de decisão das empresas.

1.1.4 Situação real em relação à informação do setor a ser implantado o software

No setor em questão, a fragmentação da informação é uma realidade presente, onde diferentes sistemas e ferramentas são utilizados para gerenciar processos específicos. Cada sistema opera de forma isolada, gerando dados que não são facilmente integráveis. Isso cria um ambiente onde a informação é dispersa e difícil de ser consolidada, prejudicando a tomada de decisões estratégicas. A falta de integração eficiente impede que a empresa tenha uma visão holística de suas operações, dificultando a identificação de padrões e tendências que poderiam ser explorados para melhorar a eficiência e a competitividade.

1.1.5 Objeto estratégico da empresa em relação ao projeto

O objeto estratégico da empresa em relação ao projeto de integração de dados é alcançar uma visão unificada e coesa das informações, permitindo uma análise mais precisa e uma tomada de decisão mais informada. Ao implementar um sistema que utiliza bancos de dados em grafos e técnicas de aprendizado de máquina, a empresa busca não apenas consolidar os dados de diferentes fontes, mas também descobrir novos insights e padrões que possam ser utilizados para otimizar processos e identificar oportunidades de negócio. Esse projeto visa transformar a maneira como a informação é gerenciada e utilizada, tornando-se um diferencial competitivo no mercado.

1.2 Descrição da solução proposta

1.2.1 Escopo do projeto

O projeto propõe o desenvolvimento de um sistema que unifique informações provenientes de diferentes fontes, como planilhas, bancos de dados e arquivos diversos, estruturando-as de forma inteligente por meio de um banco de dados em grafo. Além disso, o sistema utilizará algoritmos de Machine Learning (como regressão linear e classificação) para identificar padrões e gerar relações dinâmicas entre os dados, permitindo uma análise mais eficaz e integrada das informações empresariais.

O principal objetivo é proporcionar uma solução que facilite a integração, estruturação e análise de grandes volumes de dados heterogêneos, superando as limitações dos bancos de dados relacionais tradicionais.

1.2.2 Descrição dos usuários

Gestores e Analistas de Dados: Necessitam de uma plataforma que possibilite a unificação e análise de dados de diversas fontes, permitindo a identificação de padrões e a otimização da tomada de decisões estratégicas.

Desenvolvedores e Engenheiros de Software: Buscam uma ferramenta flexível que facilite a modelagem e organização de dados de maneira mais dinâmica do que os bancos relacionais convencionais.

Empresas e Organizações de Médio e Grande Porte: Com grande volume de dados distribuídos em diferentes sistemas, precisam de um ambiente que centralize essas informações e ofereça insights acionáveis.

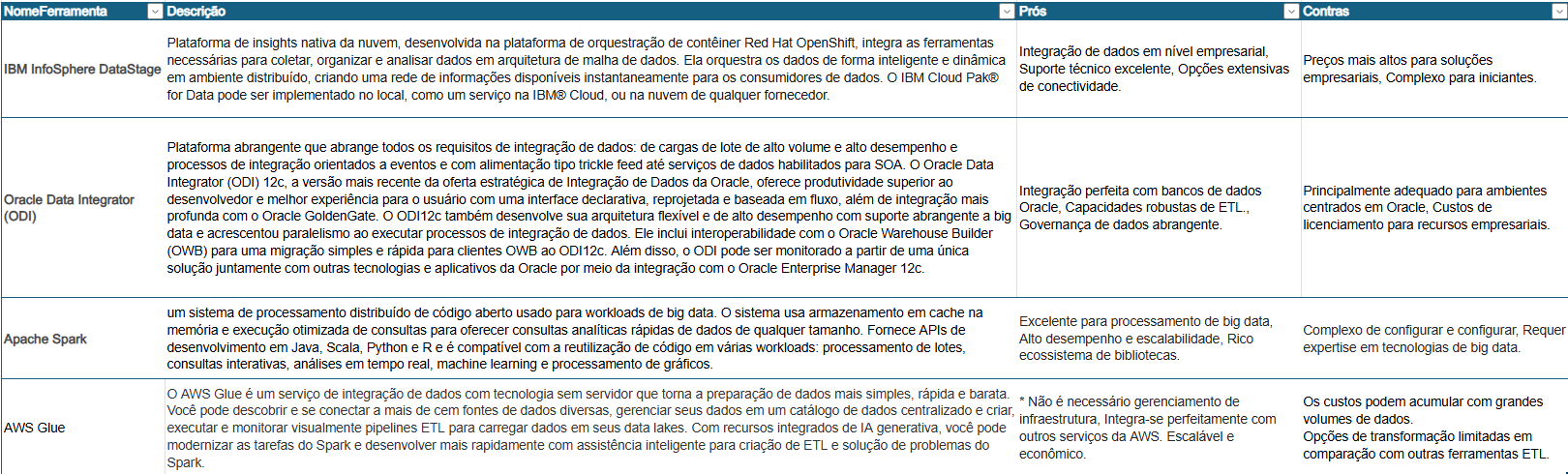
1.2.3 Descrição do ambiente

1.2.4 Perspectiva do produto em relação ao mercado

O mercado atual conta com diversas ferramentas de integração de dados, mas a maioria delas se baseia em bancos de dados relacionais, que possuem limitações quando se trata da análise de dados altamente conectados. Algumas soluções comerciais oferecem funcionalidades similares, mas poucas possuem a capacidade de criar automaticamente relações dinâmicas entre os dados usando Machine Learning, enquanto isso, as grandes empresas possuem integrações com inteligência artificial, mas o custo de aquisição ou utilização se torna muito alto.

O diferencial do projeto está na combinação entre banco de dados em grafo e algoritmos de aprendizado de máquina, permitindo não apenas integrar e armazenar os dados, mas também descobrir padrões e novas conexões automaticamente. Isso o torna uma alternativa inovadora para empresas que lidam com grande volume de dados distribuídos e necessitam de uma solução eficiente para análise e unificação dessas informações.

Com essa abordagem, o sistema pode atender nichos específicos que exigem alta flexibilidade na modelagem de dados e eficiência na recuperação de informações complexas, como empresas de tecnologia, instituições financeiras, áreas de pesquisa e análise de mercado.

1.2.5 Alternativas e concorrências

Referências

[Frontiers | Data Integration Challenges for Machine Learning in Precision Medicine](https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2021.784455/full) Acesso em 22/02

[Three Decades of Data Intecration — all Problems Solved? | SpringerLink](https://link.springer.com/chapter/10.1007/978-1-4020-8157-6_1)

Ziegler P, Dittrich KR. Three decades of data integration-All problems solved? In: *IFIP Congress Topical Sessions.* Toulouse: Springer (2004). p. 3–12.

Acesso em 22/02

[Steps Toward Large-Scale Data Integration in the Sciences: Summary of a Workshop - National Research Council, Policy and Global Affairs, Division on Engineering and Physical Sciences, Committee on Applied and Theoretical Statistics - Google Livros](https://books.google.com.br/books?hl=pt-BR&lr=&id=rCtkAgAAQBAJ&oi=fnd&pg=PP1&ots=NE8r9HpANx&sig=kG-xrvawE5e3cCaZ7YrJt96Ml7I&redir_esc=y#v=onepage&q&f=false)

Weidman S, Arrison T, et al. *Steps Toward Large-Scale Data Integration in the Sciences:: Summary of a Workshop.* Washington, DC: National Academies Press (2010).

Acesso em 22/02

[Microsoft Word - CursoJai\_final2.doc](https://www.cin.ufpe.br/~if696/referencias/integracao/JAI01.pdf)

Acesso em 25/02

[Vista do DIKW4IoT: Uma abordagem baseada na hierarquia DIKW para a construção de grafos de conhecimento para integração de dados de IoT](https://sol.sbc.org.br/index.php/sbbd/article/view/21806/21630)

Acesso em 25/02

[Knowledge graph-based data integration system for digital twins of built assets - ScienceDirect](https://www.sciencedirect.com/science/article/pii/S0926580523003692)

Acesso em 25/02

[Graph Neural Networks for Multimodal Single-Cell Data Integration | Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining](https://dl.acm.org/doi/abs/10.1145/3534678.3539213)

Acesso em 25/02

[Best Data Integration Tools Reviews 2025 | Gartner Peer Insights](https://www.gartner.com/reviews/market/data-integration-tools)

Acesso em 20/02

[Best Data Integration Tools: 26 Top Tools | Data Stack Hub](https://www.datastackhub.com/top-tools/data-integration-tools/)

Acesso em 25/02

[O que é o Spark? - Introdução ao Apache Spark e análises — AWS](https://aws.amazon.com/pt/what-is/apache-spark/)

[Oracle Data Integrator | Oracle Brasil](https://www.oracle.com/br/middleware/technologies/data-integrator.html)

[Apache Spark™ - Unified Engine for large-scale data analytics](https://spark.apache.org/)

[IBM DataStage](https://www.ibm.com/br-pt/products/datastage)

[Integração de dados com tecnologia sem servidor | AWS Glue | Amazon Web Services](https://aws.amazon.com/pt/glue/)

In today’s business world, it is typical that enterprises run different but coexisting information systems. Employing these systems, enterprises struggle to realize business opportunities in highly competitive markets. In this setting, the integration of existing information systems is becoming more and more indispensable in order to dynamically meet business and customer needs while leveraging long-term investments in existing IT infrastructure. In general, integration of multiple information systems aims at combining selected systems so that they form a unified new whole and give users the illusion of interacting with one single information system. The reason for integration is twofold: First, given a set of existing information systems, an integrated view can be created to facilitate information access and reuse through a single information access point. Second, given a certain information need, data from different complementing information systems is to be combined to gain a more comprehensive basis to satisfy the need.

//////////////////////

In this context, data integration allows the use of multiple data sources with several different (eve disparate) *pieces of evidence* to build (hopefully) interpretable models of the systems under study ([236](https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2021.784455/full#B236)). Since these broad array of data sources may have quite different structures, levels of granularity and, in the case of quantitative measurements, different distributions and dynamic ranges, data integration is indeed a demanding endeavor, briefly subsumed in the question *how can we put together these data sources to improve knowledge discovery?* ([237](https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2021.784455/full#B237)). Hence, being able to perform complex queries, build heterogeneous models and develop hierarchically nested data retrieval operations on multiple databases are core goals for data integration strategies useful for AI/ML models in Precision Medicine ([235](https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2021.784455/full#B235), [238](https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2021.784455/full#B238)–[241](https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2021.784455/full#B241)).

Uma importante decisão na construção de um sistema de integração de dados na Web é a escolha da abordagem a ser utilizada: virtual ou materializada. Na abordagem virtual, as informações são extraídas das fontes somente quando consultas são requisitadas. Por outro lado, na abordagem materializada, as informações são recuperadas, integradas e armazenadas em um repositório, de maneira que as consultas podem ser avaliadas diretamente neste repositório, sem a necessidade de acessar as fontes de dados. A Tabela I apresenta os passos que descrevem mais detalhadamente o processo de extração de informações de cada uma dessas abordagens.

Tanto a abordagem virtual quanto a abordagem materializada possui vantagens e desvantagens, sendo adequadas para situações diferentes. Se por um lado uma das principais vantagens da abordagem virtual é que as informações recuperadas estão sempre atualizadas, por outro lado, esta abordagem torna-se ineficiente quando as fontes de dados estão temporariamente inacessíveis. Sendo assim, a abordagem virtual é mais adequada para os casos em que as informações mudam rapidamente e para consultas que operam sobre uma grande quantidade de dados distribuídos em um grande número de fontes de dados. A abordagem virtual não é adequada quando os passos de tradução e integração das informações ficam muito caros ou quando as fontes de dados freqüentemente ficam inacessíveis.

No caso da abordagem materializada, sua principal vantagem é o fato de que as informações integradas estão disponíveis imediatamente para consultas, não sendo necessário acessar diretamente as fontes de dados locais. Entretanto, quando a abordagem materializada é adotada é preciso manter a consistência entre os dados armazenados no repositório de dados e os dados das fontes de dados de origem. Esta abordagem é mais adequada quando: i) são requisitadas porções específicas e previsíveis da informação disponível, ii) usuários demandam alto desempenho de consulta, sem requerer que o estado da informação seja o mais atualizado, iii) é necessário acessar cópias privadas de informação, de forma que estas possam ser modificadas, anotadas e resumidas, iv) usuários querem guardar informações que não são mantidas nas fontes de dados, tais como informações históricas. Esta abordagem não é adequada quando as informações integradas precisam estar sempre atualizadas.

Certamente, existem situações onde uma abordagem é mais adequada do que a outra. Contudo, para futuras aplicações mais complexas e em grande escala, ambas as abordagens serão necessárias. Para manipular tais aplicações, o sistema de integração de informação ideal é aquele no qual algumas informações são recuperadas, processadas e integradas previamente e, em seguida, armazenadas em um repositório do sistema, enquanto que outras informações são recuperadas e processadas somente quando consultas são requisitadas.

