ADR 1: Integração de Dados e Sistemas Baseada em Nuvem

Status
Proposto

Contexto
A BioInnovate Corp., uma empresa de biotecnologia em crescimento especializada no desenvolvimento de tratamentos para doenças crônicas, enfrenta desafios significativos na integração de dados e sistemas em seus projetos de pesquisa e inovação[1]. Atualmente, a empresa utiliza diversos sistemas como Sistemas de Gestão de Informações de Laboratório (LIMS), sistemas de gerenciamento de ensaios clínicos, bancos de dados genômicos, sistemas financeiros e de gerenciamento de projetos, e plataformas de colaboração[2]. No entanto, esses sistemas operam de forma isolada, criando silos de dados que dificultam o acesso a informações abrangentes[3].

Os principais problemas identificados incluem:

- **Silos de Dados:** Dados restritos a departamentos ou projetos, limitando a colaboração[4].
- **Formatos Inconsistentes:** Dados em formatos variados (CSV, Excel, proprietários) que exigem padronização manual[5].
- **Qualidade Deficiente:** Dados incompletos ou imprecisos comprometem a pesquisa[6].
- **Escalabilidade:** Os sistemas atuais não suportam o aumento do volume de dados[7].
- **Falta de Insights em Tempo Real:** Atrasos na integração dificultam decisões rápidas[8].
- **Conformidade:** Regulamentações como GDPR e HIPAA complicam a segurança dos dados[9].

Esses desafios resultam em atrasos nas pesquisas, limitações na colaboração interdisciplinar, desperdício de recursos e comprometimento de decisões estratégicas[10].

Decisão
A decisão tomada é implementar uma arquitetura de integração de

dados e sistemas baseada em nuvem, utilizando uma plataforma de dados unificada e serviços de integração gerenciados. O caminho escolhido envolve a migração e consolidação de dados dos diversos sistemas (LIMS, ensaios clínicos, genômicos, financeiros, etc.) para um *Data Lakehouse* na nuvem, com a implementação de pipelines de dados automatizados e ferramentas de orquestração. Além disso, será utilizada uma camada de API Gateway para facilitar a comunicação padronizada entre os sistemas e com aplicações externas.

Justificativas:

- **Silos de Dados e Formatos Inconsistentes:** Um *Data Lakehouse* permite armazenar dados em diferentes formatos (estruturados, semi-estruturados e não estruturados) de forma centralizada e escalável. Ferramentas de ETL/ELT na nuvem podem padronizar e transformar esses dados automaticamente.
- **Qualidade Deficiente:** A implementação de processos de validação e limpeza de dados nos pipelines de ingestão garantirá a qualidade dos dados antes de serem persistidos no *Data Lakehouse*.
- **Escalabilidade:** A computação em nuvem oferece escalabilidade elástica, permitindo que a infraestrutura se adapte automaticamente ao aumento do volume de dados e das demandas de processamento.
- **Falta de Insights em Tempo Real:** Pipelines de dados quase em tempo real, combinados com ferramentas de processamento de *streaming* e serviços de *analytics* na nuvem, possibilitarão insights rápidos e suporte a decisões ágeis.
- **Conformidade:** Provedores de nuvem oferecem certificações de conformidade (GDPR, HIPAA, etc.) e ferramentas de segurança robustas (criptografia, controle de acesso, auditoria) que auxiliam na aderência às regulamentações.
- **Colaboração:** A centralização dos dados e a exposição via APIs facilitam o acesso e o compartilhamento de informações entre diferentes departamentos e com parceiros.

Implicações Imediatas:

- Necessidade de equipe com conhecimento em tecnologias de nuvem e engenharia de dados.

- Custos iniciais de migração e reestruturação de dados.
- Definição de uma estratégia de governança de dados clara.

Consequências

Positivas:

- **Melhora na Colaboração:** Dados unificados e acessíveis por meio de APIs e ferramentas de análise facilitarão a colaboração interdisciplinar.
- **Decisões Baseadas em Dados:** Insights em tempo real e dados de alta qualidade permitirão decisões estratégicas mais informadas e rápidas.
- **Eficiência Operacional:** Automação da integração e processamento de dados reduzirá o esforço manual e otimizará a utilização de recursos.
- **Conformidade Reforçada:** Aderência mais fácil às regulamentações de segurança e privacidade de dados.
- **Inovação Acelerada:** A capacidade de processar e analisar grandes volumes de dados de forma eficiente pode acelerar o desenvolvimento de novos tratamentos.
- **Redução de Desperdício de Recursos:** Eliminação de duplicação de esforços na padronização e acesso a dados.

Negativas:

- **Curva de Aprendizagem:** A equipe interna pode precisar de treinamento em novas tecnologias de nuvem.
- **Custo Contínuo:** Custos operacionais contínuos associados aos serviços de nuvem (embora a longo prazo, espera-se otimização de custos).
- **Complexidade Inicial:** A migração e a configuração inicial da infraestrutura podem ser complexas.
- **Dependência do Provedor de Nuvem:** Cria uma dependência em relação ao provedor de serviços de nuvem escolhido.

Requisitos Futuros:

- Monitoramento contínuo da performance e custos da infraestrutura de nuvem.
- Implementação de uma robusta estratégia de *Data Governance*.
- Expansão dos pipelines de dados para incluir novas fontes

conforme a empresa cresce.
- Desenvolvimento de dashboards e ferramentas de BI avançadas.

Alternativas Consideradas

- **Solução On-Premise Tradicional:** Manter e expandir a infraestrutura de servidores e bancos de dados no local. Descartada devido à falta de escalabilidade, altos custos de manutenção, complexidade na gestão de diferentes formatos de dados e dificuldade em alcançar insights em tempo real de forma eficiente.
- **Integração Ponto a Ponto Manual:** Desenvolver integrações personalizadas para cada par de sistemas. Descartada por ser insustentável a longo prazo, gerando alto débito técnico, complexidade na manutenção, dificuldade em gerenciar formatos inconsistentes e falta de escalabilidade para o volume crescente de dados.

Referências

- BioInnovate Corp.pdf [11]