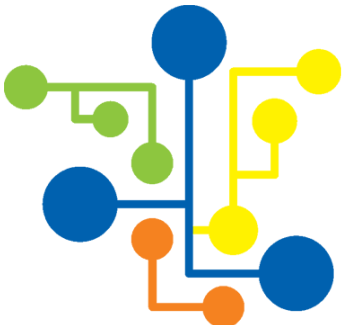


Data Science Academy

www.datascienceacademy.com.br

Data Lake – Design, Projeto e Integração

Os 4 Estágios Para Construir um Data Lake de Forma Eficiente

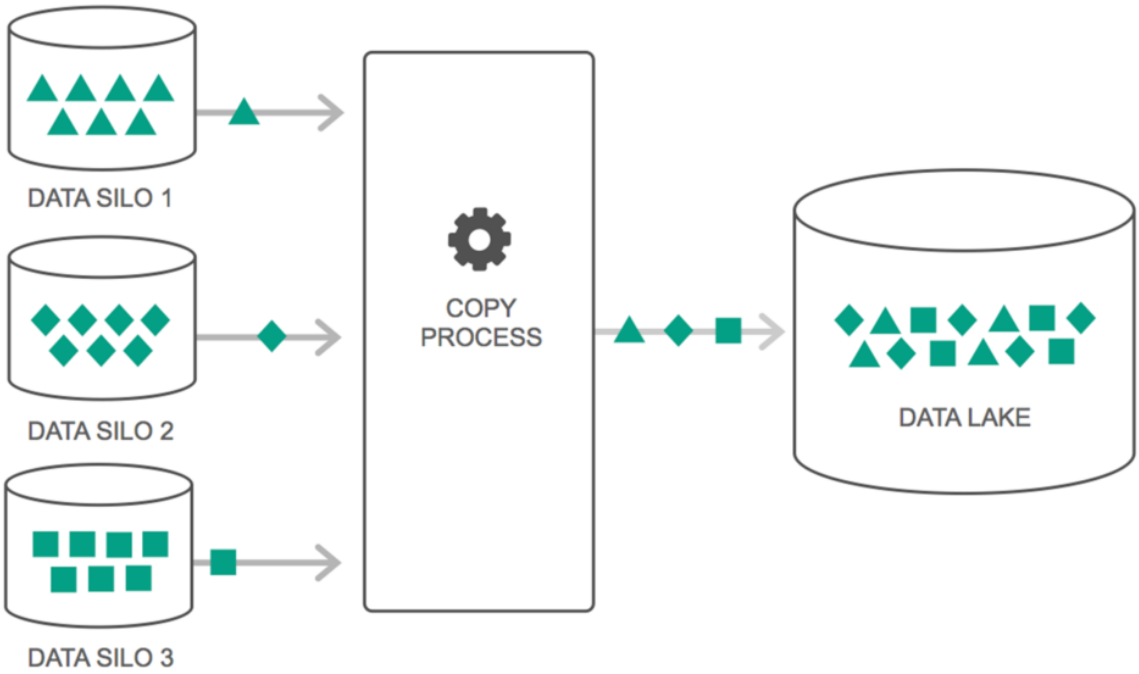
Uma abordagem ágil ao desenvolvimento de um Data Lake pode ajudar as empresas a lançar programas analíticos rapidamente e estabelecer uma cultura orientada a dados. A consultoria McKinsey traçou um raio-x sobre Os 4 Estágios Para Construir um Data Lake de Forma Eficiente, o qual publicamos aqui para você na íntegra e em português.

Nota: este material está disponível gratuitamente em nosso blog, mas você recebe aqui com exclusividade a versão em pdf. Boa leitura!

O aumento da capacidade de processamento de computadores, capacidade e uso de armazenamento em nuvem e conectividade de rede estão transformando o atual fluxo de dados na maioria das empresas em um maremoto - um fluxo infinito de informações detalhadas sobre os perfis pessoais dos clientes, dados de vendas, especificações de produtos, etapas de processos e assim por diante. Os dados chegam em todos os formatos e em uma variedade de fontes, incluindo dispositivos de Internet-of-Things, sites de mídia social, sistemas de vendas e sistemas de colaboração interna. Big Data na sua essência.

Apesar do aumento no número de ferramentas e tecnologias projetadas para facilitar a coleta, o armazenamento e a avaliação de informações críticas de negócios, muitas empresas ainda não sabem ao certo como lidar com esses dados. Os líderes de negócios e TI continuam sobrecarregados pelo grande volume e variedade de dados à sua disposição, a velocidade com que as informações estão atravessando as redes internas e externas e o custo de gerenciar toda essa inteligência de negócios. Cada vez mais, eles também estão sendo encarregados de uma tarefa ainda mais complicada: aproveitar percepções significativas de todas essas informações comerciais.

Esses executivos devem expandir suas infraestruturas de gerenciamento de dados de forma massiva e rápida. Uma classe emergente de tecnologias de gerenciamento de dados tem uma promessa significativa a esse respeito: os Data Lakes. Essas plataformas de armazenamento são projetadas para armazenar, processar e analisar dados estruturados e não estruturados. Eles são normalmente usados em conjunto com os Data Warehouses corporativos tradicionais (EDWs - Enterprise Data Warehouses), mas, em geral, custam menos para operar do que os EDWs. O resultado da economia de custos se deve ao fato que as empresas podem usar hardware acessível e fácil de obter e porque os conjuntos de dados não precisam ser indexados e preparados para armazenamento no momento da ingestão. Os dados são mantidos em seus formatos nativos e reconfigurados apenas quando necessário, conforme necessário. Os bancos de dados relacionais também podem precisar ser gerenciados como parte da plataforma de banco de dados, mas apenas para facilitar a capacidade dos usuários finais de acessar algumas fontes de dados.



Como os dados são carregados em formatos "brutos" em vez de pré-configurados quando entram nos sistemas da empresa, eles podem ser usados de maneiras que vão além da captura básica. Por exemplo, os Cientistas de Dados que podem não saber exatamente o que estão procurando podem encontrar e acessar dados rapidamente, independentemente do formato. De fato, uma “zona de dados brutos” bem mantida e governada pode ser uma mina de ouro para Cientistas de Dados que buscam estabelecer um robusto programa de análise avançada. E à medida que as empresas ampliam seu uso de Data Lakes além de apenas pequenos projetos-piloto, elas podem estabelecer opções de "autoatendimento" para usuários corporativos, nos quais podem gerar suas próprias análises e relatórios de dados.

No entanto, pode ser demorado e complicado integrar Data Lakes a outros elementos da arquitetura de tecnologia, estabelecer regras apropriadas para o uso de Data Lakes em toda a empresa e identificar os produtos, talentos e recursos de suporte necessários para implantar Data Lakes e realizar benefícios significativos para os negócios. Por exemplo, as empresas normalmente não têm experiência em certas abordagens de gerenciamento de dados e precisam encontrar funcionários fluentes em tecnologias emergentes de Streaming, como Flume e Spark. Encontrar Engenheiros de Dados que dominem estas tecnologias e saibam implementar Data Lakes de forma profissional e eficiente é um grande desafio em um mercado com alta carência de pessoas com esses perfis.

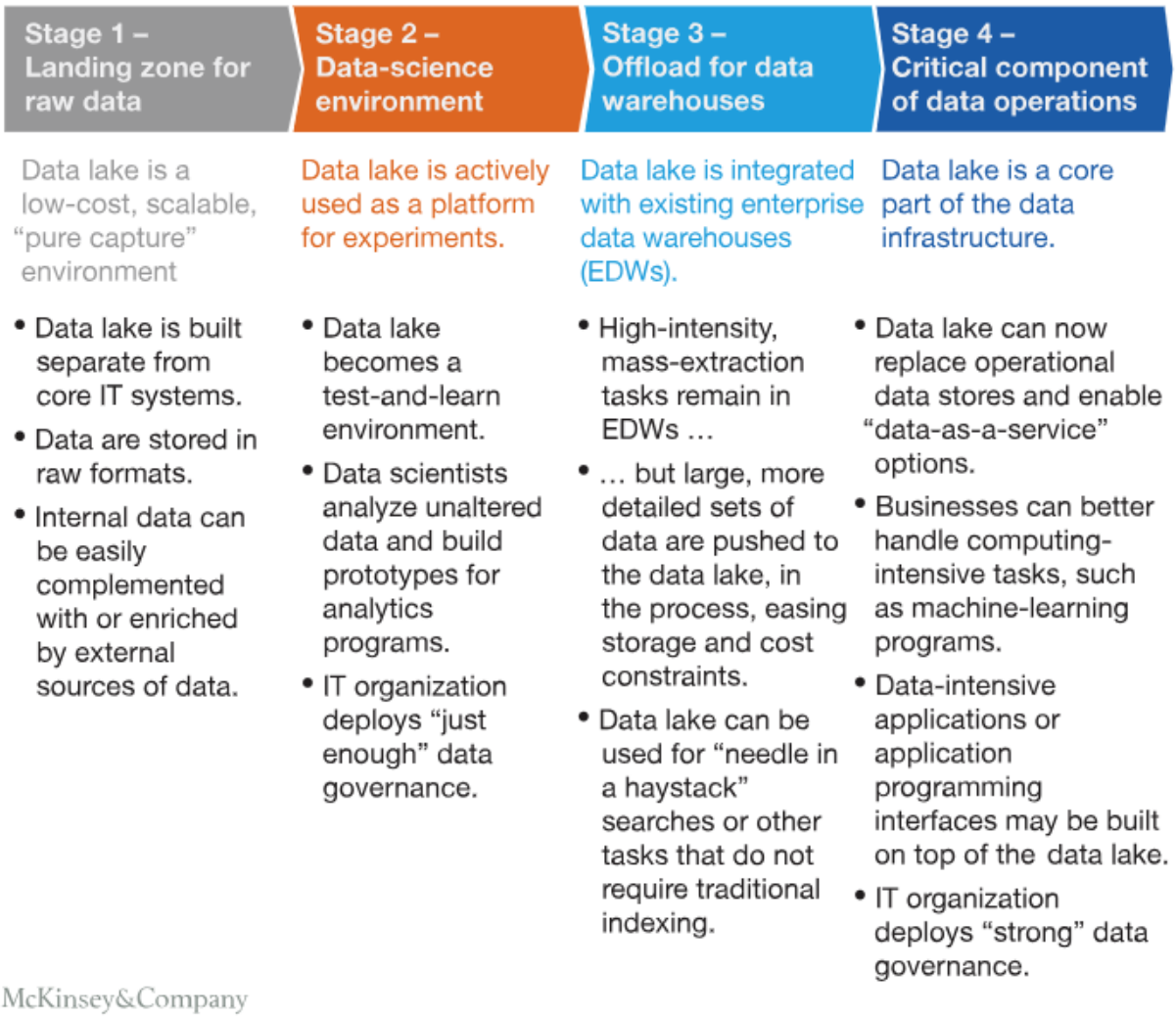
As empresas estão recorrendo a métodos testados e comprovados para atualizar arquiteturas de tecnologia - por exemplo, participando de longas discussões internas sobre designs, produtos e fornecedores ideais e evitando a construção de uma solução de

armazenamento até que tenham bem claro os problemas que podem ser resolvidos com a solução adotada.

As empresas devem aplicar uma abordagem ágil ao design e à implementação de Data Lakes, testando uma série de tecnologias e abordagens de gerenciamento e testando-as e refinando-as antes de chegar aos processos ideais de armazenamento e acesso de dados. As empresas que o fazem conseguem acompanhar as rápidas mudanças nos padrões regulatórios e de conformidade dos dados - por exemplo, o Regulamento Geral de Proteção de Dados da União Europeia (GDPR), que entrou em vigor em maio de 2018. Talvez mais importante, as empresas podem trazer percepções orientadas por análises de mercado muito mais rápido do que seus concorrentes, reduzindo significativamente o custo e a complexidade de gerenciar sua arquitetura de dados.

Estágios do Desenvolvimento de um Data Lake

As empresas geralmente passam pelos quatro estágios seguintes de desenvolvimento ao construir e integrar Data Lakes dentro de suas arquiteturas de tecnologia existentes:



Vejamos cada um desses estágios:

Estágio 1 - Landing zone or raw data. No primeiro nível, o Data Lake é construído separado dos principais sistemas de TI e serve como um ambiente de "captura pura" de baixo custo e escalável. O Data Lake serve como uma fina camada de gerenciamento de dados dentro da pilha de tecnologia da empresa que permite que dados brutos sejam armazenados indefinidamente antes de serem preparados para uso em ambientes de computação. As organizações podem implantar o Data Lake com efeitos mínimos na arquitetura existente. Uma

governança forte, incluindo classificação rigorosa de dados, é necessária durante essa fase inicial se as empresas desejarem evitar a criação de um pântano de dados (Data Swamp).

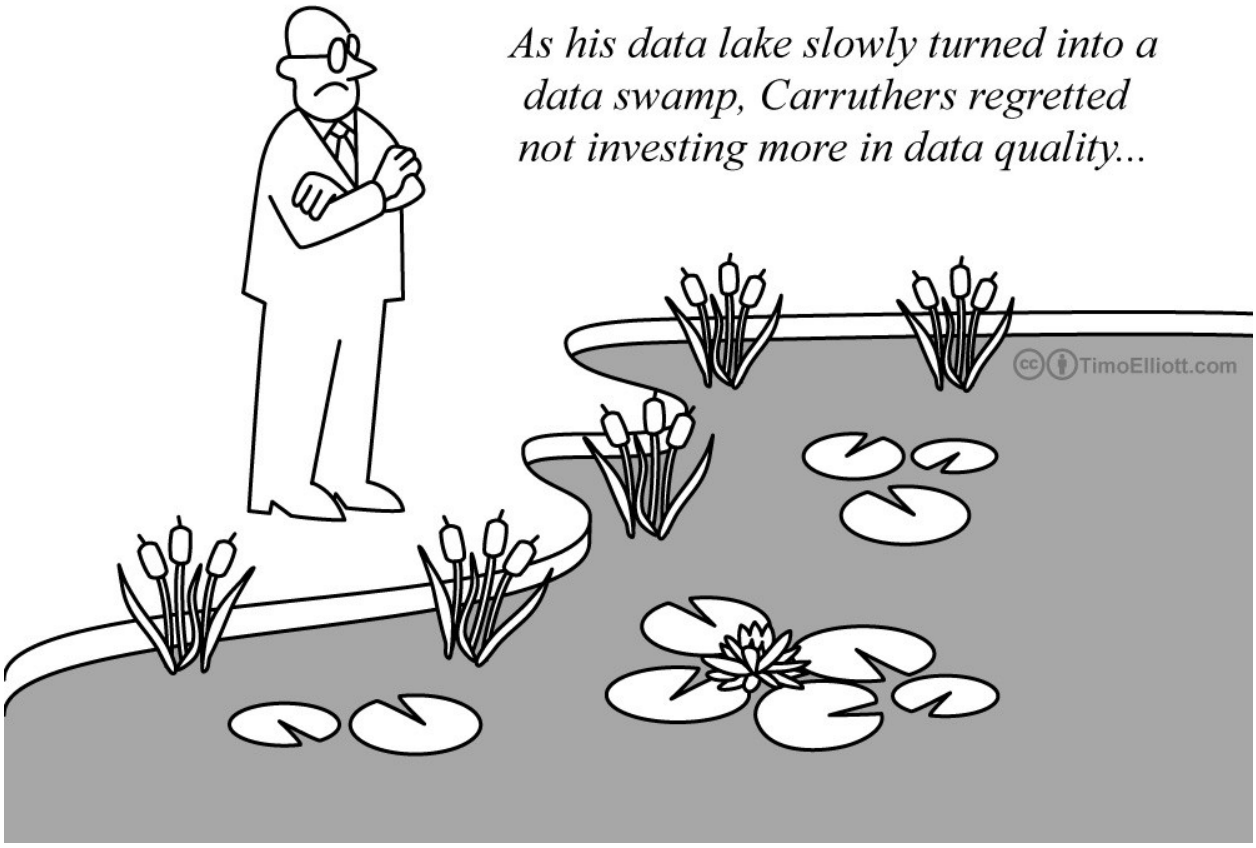
Estágio 2 - Data Science Environment. Neste próximo nível, as organizações podem começar a usar mais ativamente o Data Lake como uma plataforma para experimentação. Os Cientistas de Dados têm acesso fácil e rápido aos dados - e podem se concentrar mais na execução de experimentos com dados e na análise de dados, em vez de se concentrarem apenas na coleta e aquisição de dados. Neste "sandbox", eles podem trabalhar com dados inalterados para criar protótipos para programas analíticos. Eles podem implantar uma variedade de ferramentas comerciais e de código aberto ao lado do Data Lake para criar os ambientes de teste necessários.

Estágio 3 - Offload for Data Warehouses. No próximo nível, os Data Lakes estão começando a ser integrados aos EDWs existentes. Aproveitando os baixos custos de armazenamento associados a um Data Lake, as empresas podem armazenar dados "frios" (raramente usados ou inativos). Elas podem usar esses dados para gerar insights sem pressionar ou exceder as limitações de armazenamento, ou sem precisar aumentar drasticamente o tamanho dos Data Warehouses tradicionais. Enquanto isso, as empresas podem manter a extração de dados relacionais de alta intensidade em EDWs existentes, que têm o poder de lidar com eles. Elas podem migrar tarefas de extração e transformação de baixa intensidade para o Data Lake - por exemplo, uma pesquisa do tipo "agulha no palheiro", na qual os Cientistas de Dados precisam varrer bancos de dados para consultas não suportadas por estruturas de índices tradicionais.

Estágio 4 - Critical component of data operations. Uma vez que as empresas cheguem a esse estágio de lançamento e desenvolvimento, é muito provável que grande parte das informações que circulam pela empresa estejam passando pelo Data Lake. O Data Lake se torna uma parte essencial da infraestrutura de dados, substituindo Data Marts existentes ou armazenamentos de dados operacionais e permitindo o fornecimento de dados como um serviço. As empresas podem aproveitar ao máximo a natureza distribuída da tecnologia de Data Lake, bem como sua capacidade de lidar com tarefas de uso intensivo de computação, como aquelas exigidas para conduzir análises avançadas ou implantar programas de aprendizado de máquina (Machine Learning). Algumas empresas podem decidir criar aplicativos com uso intensivo de dados na parte superior do Data Lake - por exemplo, um painel de gerenciamento de desempenho. Ou elas podem implementar interfaces de programação de aplicativos para que possam combinar perfeitamente os insights obtidos dos recursos do Data Lake com insights obtidos de outros aplicativos.

O tempo e os recursos necessários para que as empresas ampliem seus Data Lakes, desde simples zonas de pouso até componentes críticos da infraestrutura de dados, variam de acordo com os objetivos e pontos de partida das empresas. Em cada estágio do desenvolvimento, as empresas precisam examinar questões complicadas relacionadas ao tamanho e à variedade de seus conjuntos de dados, seus recursos existentes no gerenciamento

de dados, o nível de especialização em Big Data em suas unidades de negócios e conhecimento do produto na organização de TI. Por exemplo, quão sofisticadas são as ferramentas de análise no ambiente atual? A empresa está usando ferramentas e metodologias de desenvolvimento tradicionais ou mais novas? Quantos usuários de dados concorrentes a empresa normalmente exige? As cargas de trabalho são gerenciadas dinamicamente? Com que rapidez os usuários finais precisam acessar os dados? Em vários pontos do processo de desenvolvimento do Data Lake, as empresas podem ficar atoladas nesses detalhes e perder o ímpeto.



A jornada do Data Lake de "projeto científico" para o componente totalmente integrado da infraestrutura de dados pode ser acelerada, no entanto, quando os líderes de TI e negócios se unem para responder a essas e outras perguntas sob um modelo de desenvolvimento ágil. Em nossa experiência, uma abordagem ágil pode ajudar as empresas a obter vantagens de seus Data Lakes em meses, em vez de anos. Vitórias rápidas e evidências de impacto de curto prazo podem ajudar muito a manter os líderes de TI e negócios envolvidos e concentrados em questões de gerenciamento de dados - limitando assim a necessidade de retrabalho futuro e ajustes intermináveis de protocolos associados ao preenchimento, gerenciamento e acesso ao Data Lake. Uma abordagem ágil pode colocar os líderes de TI e de negócios na mesma sintonia.

Essa colaboração é fundamental não apenas para determinar um caminho técnico para o Data Lake, mas também para estabelecer um ambiente de trabalho compatível com dados e aproveitar novas oportunidades de negócios com base em insights preciosos.

Construindo um Data Lake: Uma Abordagem Ágil

A maioria das organizações compreende a necessidade de metodologias ágeis no contexto do desenvolvimento de software e agora precisam aplicar essas metodologias no contexto do gerenciamento de dados. Normalmente, a organização de TI toma a iniciativa de examinar as possíveis opções de tecnologia e abordagens para a construção de Data Lakes, com poucas informações das unidades de negócios. Sob uma abordagem ágil, os líderes de TI e de negócios definem e abordam questões relevantes sobre tecnologia e design. Por exemplo, o Data Lake será construído usando uma solução on-premises ou será hospedado na nuvem (usando servidores externos, públicos ou híbridos)? Como o Data Lake será preenchido - isto é, quais conjuntos de dados fluirão para o lago e quando? Idealmente, a população do Data Lake deve se basear nos usos de negócios de prioridade mais alta e ser feita em ondas, em oposição a um grande esforço de uma só vez para conectar todos os fluxos de dados relevantes dentro do Data Lake.

De fato, os adotantes iniciais mais bem-sucedidos estão projetando seus Data Lakes usando uma abordagem de "businessback", em vez de considerar fatores de tecnologia primeiro. Eles estão identificando os cenários em que as unidades de negócios poderiam obter o máximo valor do Data Lake e, em seguida, incluir esses cenários no design (ou reprojeto) da solução de armazenamento e nas decisões de rollout. Em seguida, as empresas estão preenchendo o Data Lake com dados para grupos específicos ou casos de uso, conforme necessário. E, em vez de incluir tudo em uma solução designada, as empresas estão testando dois ou três candidatos finais de diferentes fornecedores para avaliar o desempenho no mundo real, a facilidade de integração e a escalabilidade de suas ofertas.

Essa abordagem ágil de implantação pode garantir que os desafios de desempenho ou implementação sejam detectados antecipadamente e incorpore o feedback das unidades de negócios. Também deixa espaço para que as equipes de desenvolvimento ágil modifiquem os processos e os protocolos de controle de dados à medida que o Data Lake se enche, as tecnologias de análise e armazenamento mudam e os requisitos de negócios evoluem.

À medida que os Data Lake passam de projetos piloto para elementos centrais da arquitetura de dados, os líderes de negócios e de tecnologia precisarão reconsiderar suas estratégias de governança. Especificamente, eles precisam aprender a equilibrar a rigidez da supervisão tradicional de dados com a necessidade de flexibilidade, à medida que os dados são rapidamente coletados e usados em um mundo digital. Sob uma abordagem ágil de governança, as empresas podem aplicar supervisão suficiente à medida que novas fontes entram no Data Lake, evitando algumas das práticas de engenharia mais rígidas exigidas nos

Data Warehouses tradicionais e refinando regras e processos conforme os requisitos de negócios determinam para chegar a uma solução ideal. Por exemplo, os Cientistas de Dados podem receber sinal livre para explorar dados, mesmo que os casos de negócios para certas categorias de dados ainda estejam sendo identificados. Enquanto isso, os usuários finais podem enfrentar controles mais rigorosos até que os casos de uso sejam estabelecidos com mais firmeza.

No mínimo, no entanto, as empresas devem designar certos indivíduos como proprietários de conjuntos de dados e processos, para que as responsabilidades sejam claras e as decisões sobre fontes de dados e direitos de acesso possam ser feitas rapidamente. Como os dados não estão sendo estruturados com antecedência, as empresas também desejam capturar e armazenar metadados em todas as fontes de dados que fluem para o lago (dentro do próprio lago ou em um registro separado) e manter um catálogo de dados central para todas as partes interessadas. Além disso, as empresas podem precisar reconfigurar os direitos de acesso à medida que se repetem nos protocolos de gerenciamento de dados, tendo em mente os requisitos regulamentares e os problemas de privacidade relacionados à retenção de informações pessoalmente identificáveis. Os proprietários de dados devem comunicar esses direitos de acesso a todas as partes interessadas relevantes.

Caso de Uso: Transformação Nos Serviços Bancários com Data Lake

Vamos considerar como um banco global aplicou princípios ágeis ao desenvolvimento de um Data Lake. O banco vinha enfrentando vários desafios críticos de dados: informações de negócios de baixa qualidade, falta de especialistas para gerenciar diferentes conjuntos de dados que chegam em diferentes formatos, tecnologias antigas de Data Warehouse e mais de 1.000 fontes de dados. Os sistemas eram desajeitados. Os conjuntos de dados de entrada precisavam ser estruturados antes que pudessem ser inseridos em camadas no Data Warehouse e antes que quaisquer relatórios utilizáveis pudessem ser criados.

Além desses desafios técnicos, os líderes de negócios e de TI do banco não estavam trabalhando de maneira colaborativa, o que exacerbou os problemas de dados da empresa. Os dados estavam sendo armazenados em sistemas isolados, de modo que informações críticas de negócios geralmente ficavam presas e as solicitações de acesso a determinados conjuntos de dados demoravam a obter uma resposta devido à má coordenação e comunicação entre unidades de negócios e operações de TI. O gerenciamento de dados era visto como "trabalho de TI". Os líderes de negócios mantinham o assunto à distância e, portanto, lutavam para articular suas necessidades de dados.

Os líderes seniores do banco estavam preocupados com a perda de clientes, em parte devido à incapacidade da empresa de gerenciar os dados com habilidade. Eles decidiram experimentar tecnologias de Data Lakes para tentar facilitar a extração, estruturação e entrega de conjuntos de dados. Buscando trabalhar tão rapidamente quanto seus desenvolvedores de

software, a empresa usou um modelo de desenvolvimento ágil e implementou o projeto de Data Lake em fases.

Os líderes seniores reuniram uma equipe de dados ágil envolvendo especialistas no assunto das unidades de negócios e da organização de TI para considerar o impacto nos negócios e os casos de uso para melhorar a qualidade e o acesso aos dados antes de determinar quais áreas da empresa teriam acesso inicial ao Data Lake.

A equipe de dados ágil conduziu entrevistas detalhadas com usuários corporativos para identificar pontos problemáticos e oportunidades nas práticas existentes de gerenciamento de dados. O plano da equipe era liberar ondas de novos serviços de dados e aplicativos em janelas de quatro meses - implementando novas ferramentas de gerenciamento de dados, desenvolvendo serviços de entrega de dados com as unidades de negócios e refinando processos com base no feedback dos clientes. Em alguns meses após o lançamento inicial do projeto de dados ágil, o banco conseguiu carregar dados relevantes para casos de uso de negócios específicos em um ambiente comum e identificar os elementos de dados críticos necessários para fornecer serviços às unidades de negócios.

O sucesso em áreas de alto perfil do negócio permitiu que o banco estendesse o uso do Data Lake para outras áreas nos meses subsequentes. A mudança da estruturação de todos os dados para a documentação de um processo backend apenas para os dados utilizados foi significativa. O banco conseguiu decompor os silos de dados; Agora, as informações dos sistemas podiam ser encontradas em um único lugar, e os funcionários podiam acessar várias formas de dados (demográficas, geográficas, mídias sociais e assim por diante) para obter uma visão de 360 graus dos clientes. A colaboração entre as unidades de negócios e o grupo de TI aumentou, assim como as pontuações de satisfação dos funcionários e dos clientes.

Mais e mais empresas estão experimentando Data Lakes, na esperança de capturar vantagens inerentes em fluxos de informação que são prontamente acessíveis, independentemente da plataforma e do business case, e que custam menos para armazenar do que os dados em armazéns tradicionais. No entanto, como acontece com qualquer implantação de novas tecnologias, as empresas precisarão reimaginar sistemas, processos e modelos de governança. Haverá questões inevitáveis sobre protocolos de segurança, pools de talentos e a construção de arquitetura empresarial que garanta flexibilidade não apenas dentro das pilhas de tecnologia, mas também nos recursos de negócios. Nossa experiência sugere que uma abordagem ágil para a implementação de Data Lakes pode ajudar as empresas a aumentar a curva de aprendizado de maneira rápida e eficaz.

Traduzido do artigo original: A smarter way to jump into data lakes da Consultoria McKinsey.