



CONSULTANCY

Modernização de Arquiteturas de Dados para uma Era Digital usando a Virtualização de Dados

Um Whitepaper

Rick F. van der Lans
Analista de Business Intelligence Independente
R20 / Consultoria

Outubro 2019

Patrocinado por

denodo 

Copyright © 2019 R20 / Consultoria. Todos os direitos reservados. Denodo Technologies e Denodo são marcas registradas da Denodo Technologies e outros nomes de produtos, nomes de serviços, slogans ou logotipos da Denodo mencionados neste documento são marcas comerciais ou marcas comerciais registradas da Denodo Technologies. Todos os outros nomes de empresas e produtos mencionados neste documento podem ser marcas comerciais ou marcas registradas de seus respectivos proprietários.

Índice

	Resumo executivo	1
1	O Novo Mundo da Entrega de Dados	2
2	Os Desafios de Criar Novas Arquiteturas	3
3	A Virtualização de Dados como Catalisador para Arquiteturas de Dados Modernas	6
4	Outras Soluções para Modernização da Arquitetura de Dados	11
5	Comentários finais	13
	Sobre o autor	15
	Sobre Denodo	15

Resumo executivo

Introdução — Inúmeras organizações começaram a fazer *projetos de modernização de arquiteturas de dados*. Isso porque as empresas começam a entender o crescente valor comercial dos dados. Ao usar dados de maneira mais eficaz, ampla e profunda, é possível melhorar e potencializar os processos de negócios e de tomada de decisão. É, na verdade, um requisito para se manter competitivo, econômico e eficiente. Basicamente, as organizações querem e precisam tornar-se mais orientadas a dados para participar efetivamente da economia digital emergente.

Arquiteturas de dados atuais não são suficientes e precisam ser modernizadas.

Elas entendem o potencial valor comercial dos dados, mas também entendem que a arquitetura dos sistemas de entrega existentes, como o *data warehouse* (e até certo ponto seus *data lakes*), não são suficientes. Esses sistemas que dependem da movimentação física de dados e o armazenamento redundante de dados nem sempre têm o desempenho, a escalabilidade e a funcionalidade corretos. Para que as arquiteturas de dados estejam à prova de futuras evoluções da análise de sistemas, as atuais precisam ser modernizadas.

Novas Necessidades de Negócios — Existe uma grande variedade de razões pelas quais as empresas precisam se modernizar, como por exemplo permitir o auto-atendimento na geração de relatórios e interfaces; garantir que as análises avançadas possam trabalhar com dados (quase) em tempo real, em vez dos dados de ontem; combinar dados internos com dados externos provenientes de uma das muitas fontes públicas, comerciais ou sociais para enriquecer insights analíticos; implantar IA / *machine learning* / ciência de dados para descobrir padrões ou tendências nos dados que possam melhorar decisões, automatizar ou otimizar processos de negócios; e implantar a tecnologia *IoT* para monitorar máquinas e processos de negócios com muito mais detalhes para melhorar a eficiência e reduzir os riscos.

Além disso, outros tipos de requisitos também têm impacto nos sistemas de BI, como o efeito de novos regulamentos sobre proteção e privacidade de dados. Nem todo mundo tem autorização para ver todas as informações de identificação pessoal (PII, na sigla em inglês); alguns dados precisam ser *anonimizados* e 'o direito de ser esquecido' precisa ser implementado de alguma forma. Mais usuários e reguladores de negócios exigem que toda a produção de dados se torne mais transparente, o que implica que catálogos e meta-dados precisem estar mais atualizados.

Modernização e não Substituição — A modernização normalmente não é apenas uma questão de usar mais computadores, substituir uma ferramenta por outra, substituir o servidor de banco de dados atual por um mais rápido ou migrar dados para a nuvem. Não há solução rápida. E substituir a arquitetura inteira por uma nova também não é uma alternativa viável. Isso seria muito arriscado. Para que a modernização seja eficaz, ela deve ser contínua. As operações comerciais atuais não podem parar por causa dessa mudança. A modernização de uma arquitetura de dados não é uma simples substituição de uma arquitetura por outra, mas envolve a melhoria dos módulos existentes e a remoção de módulos fracos.

A modernização das arquiteturas de dados deve ser um processo contínuo.

O Whitepaper — Este *whitepaper* descreve como a *virtualização de dados* pode ajudar a ter uma evolução contínua dos recursos de uma arquitetura de dados existente. Com a virtualização, as arquiteturas de dados podem ser modernizadas sem atrapalhar a carga de trabalho analítico existente. Basicamente, a virtualização de dados pode ampliar uma arquitetura atual para desbloquear e explorar mais rapidamente todos os dados existentes, apresentar mais dados de baixa latência e apoiar novas formas de uso, como a ciência de dados, sem a necessidade de substituição total.

A virtualização de dados pode ajudar a modernizar arquiteturas de dados sem a necessidade de uma substituição completa.

O *data lake*, o auto-atendimento de BI, a tecnologia em nuvem e o catálogo de dados são frequentemente mencionados em projetos de modernização. Esse *whitepaper* também descreve como a virtualização de dados pode ajudar a simplificar a inclusão de tais conceitos em uma arquitetura de dados nova e *à prova de futuro*.

1 O Novo Mundo da Entrega de Dados

Os dados se tornaram um ativo comercial essencial — Todos tem notado que o papel dos dados dentro das organizações cresceu. Os principais executivos, especialmente, entendem o potencial valor comercial dos dados. Durante muito tempo, armazenar e processar dados foi visto como uma necessidade. Os dados precisavam ser armazenados para dar suporte aos principais processos de negócios, como pedidos de cobranças, ordens de pagamento e planejamento de estoque. Mas muito mais pode ser feito com dados. Novas tecnologias e ferramentas nos permitem ser mais criativos. Quando usados da maneira correta, podem fortalecer a posição competitiva da empresa, aumentar sua participação de mercado, otimizar e potencializar processos de negócios, minimizar o desperdício industrial, melhorar o transporte de mercadorias, implantar recursos com mais eficiência, verificar automaticamente transações financeiras duvidosas e assim por diante.

Demorou algum tempo, mas coisas como *transformação digital*, querer ser uma *organização orientada por insights* e exploração da IA e de *machine learning* convenceram os principais executivos de que os dados são um ativo essencial dos negócios. Usar dados para ganhar uma vantagem competitiva tornou-se um tópico recorrente nas reuniões de conselhos de administração.

Elevando os Padrões dos Sistemas de Entrega de Dados — A crescente importância dos dados faz com que seja necessário elevar os padrões dos sistemas de análise que suportam os processos de negócios e de tomada de decisão. As organizações querem fazer mais com os dados que possuem, explorando-os de maneira mais eficaz, eficiente e ampla. O que significa ‘fazer mais’ com dados?

As novas demandas têm exigido mais dos sistemas de análise.

- Permitir o auto-atendimento na geração de relatórios e acesso a interfaces; para a análise de dados funcionar, tem que ser (quase) em tempo real. Apresentar os dados de ontem não é mais aceitável. Oferecer interfaces de serviço nas quais terceiros possam solicitar dados sem intervenção humana.
- Combinar dados internos com externos provenientes de uma das muitas fontes de dados públicas, comerciais ou sociais para enriquecer os recursos analíticos.
- Acelerar iniciativas de IA / *machine learning* onde cientistas de dados possam descobrir padrões ou tendências que antes eram desconhecidos. Isso pode levar ao desenvolvimento de modelos preditivos ou descritivos que ajudam a melhorar as decisões ou otimizar os processos de negócios.
- Simplificar a implantação da tecnologia de IoT para gerar mais dados nas máquinas e nos processos de negócios. Em outras palavras, criar dados mais detalhados para, possivelmente, ter mais *insights*.
- Oferecer análise de dados *edge*, na qual dados são estudados continuamente e em tempo real, e próximos de onde são produzidos pelo sensor ou processo de negócios (o *edge*).

Outros tipos de requisitos também afetam os sistemas de análise e entrega de dados: regulamentações novas, como a LGPD, e a exigência dos usuários e reguladores de que toda a “produção” de dados seja mais transparente. A transparência significa que a empresa deve mostrar como os dados são transportados do começo ao fim, desde a fonte até os relatórios, e como são filtrados, limpos, transformados, integrados e agregados no processo.

Atualizando os Sistemas Analíticos — Colocar dados como ponto focal de uma organização tem um enorme impacto nos sistemas de análise de dados existentes, como os de BI, de geração de relatórios, de entrada de dados e sites. A quantidade de dados armazenados será maior, o nível de detalhamento também, e mais relatórios serão processados. O número de usuários que tem acesso aos sistemas também crescerá, assim como a forma que usam os dados e a precisão exigida. Além disso, como os dados fazem parte de muitos processos de decisão, a qualidade dessas decisões depende diretamente da qualidade dos dados. Dados incorretos ou obsoletos levam a decisões incorretas.

As organizações têm dependido cada vez mais da disponibilidade e qualidade dos dados

Em outras palavras, como os padrões de uso de dados estarão mais altos, as empresas dependem cada vez mais de sua disponibilidade e qualidade, e isso afeta os sistemas de análise. No entanto, a implementação de todos os requisitos acima pode atrapalhar os sistemas existentes, porque ultrapassa seus limites técnicos. Isto leva a:

- fraco desempenho
- demora para apresentar novos relatórios
- instabilidades do sistema
- oportunidades de negócios perdidas
- desenvolvimento de relatórios isolados nas departamentos

A principal razão para essa interrupção é que as arquiteturas de dados, ferramentas de integração, aplicativos e processos existentes não foram inicialmente selecionados e projetados para esse aumento na carga de trabalho. Elas já estão trabalhando além de seus limites.

Os sistemas existentes foram esticados além de seus limites.

Modernizando a Arquitetura de Dados — A modernização normalmente não é apenas uma questão de usar mais computadores, de substituir uma ferramenta por outra, de substituir o servidor de banco de dados atual por um mais rápido ou migrar dados para a nuvem. Não há solução rápida. Muitas vezes, toda a arquitetura responsável pela entrega de dados deve ser reavaliada e modernizada. Mudanças profundas precisam ser feitas. Os fluxos de dados podem precisar ser redesenhados, novos repositórios de dados desenvolvidos, a terceirização de todo o sistema para uma plataforma em nuvem que possa ser redimensionada pode ser a solução, assim como o uso de outras técnicas para projetar estruturas de dados, e assim por diante.

Mas independentemente do que precisa ser feito para dar suporte à nova carga de trabalho, substituir a arquitetura inteira por uma nova quase nunca é a melhor opção, porque é uma estratégia considerada arriscada. Se for para modernizar, que isso seja feito de forma contínua e sem interrupções. A carga de trabalho atual não pode e não deve sofrer durante esse exercício. A modernização da arquitetura de dados não é uma simples substituição de uma arquitetura por uma nova. A modernização envolve a melhoria dos módulos existentes e a remoção de módulos fracos.

A modernização envolve a melhoria dos módulos existentes e a remoção de módulos fracos.

2 Os Desafios de Criar Novas Arquiteturas

A seção anterior mencionou vários motivos comerciais para a modernização das arquiteturas de dados. Esta seção descreve alguns dos desafios técnicos resultantes das razões comerciais que estão por trás do desenvolvimento de novas arquiteturas para sistemas de entrega de dados.

Dados em Tempo Real — Os usuários corporativos precisam cada vez mais acessar dados em tempo quase real. Tomadores de decisões operacionais, em especial, a própria força de trabalho e os clientes on-line não podem trabalhar com os dados de desatualizados.

Além disso, para muitas decisões automatizadas *dados em tempo real* (ou dados com latência zero) são uma necessidade. Os dados devem ser analisados logo após serem produzidos. A necessidade de ter dados em tempo real se aplica, por exemplo, a vários ambientes industriais, sistemas de jogos e sistemas financeiros.

Dados em tempo real são necessários para muitas decisões

Tecnicamente, isso significa que os dados devem ser levados aos consumidores de dados assim que forem produzidos. A maioria dos sistemas existentes não foi projetada para uso de dados em tempo real e por isso não tem a tecnologia adequada. A maneira como eles são projetados agora, copiando os dados várias vezes de um banco de dados para outro usando vários processos ETL, não funciona nesse caso. Cria muita latência de dados e potencialmente introduz erros no processo. Uma nova arquitetura de dados deve ser capaz de suportar a funcionalidade de transmitir dados para consumidores ou de reproduzi-los em tempo real.

Preservando o Histórico de Dados – Antigamente, o armazenamento em disco era caro. Os engenheiros faziam tudo para projetar sistemas de maneira que a quantidade de dados armazenados fosse mínima. Um dos truques predominantes usados para minimizar o armazenamento de dados foi limitar o histórico de dados. A maioria das alterações nos dados não era guardada. Valores e registros antigos eram simplesmente excluídos e substituídos por novos. Às vezes, nenhum histórico era preservado ou apenas uma porção limitada era mantida.

A remoção de dados economiza espaço em disco, mas também limita os recursos analíticos, especialmente aqueles em que os processos precisam ser analisados historicamente. Um requisito para a organização pode ser armazenar um histórico completo da maioria dos dados. Isso pode ter uma consequência séria para as ferramentas e tecnologias necessárias em uma nova arquitetura e na forma como mantemos o controle de todos esses dados.

As organizações podem precisar do histórico completo dos dados.

Big Data – Este está relacionado ao tópico anterior e ainda é bastante popular no setor de TI, o famoso *big data*. Observe que esse termo está começando a desaparecer, porque grandes volumes de dados estão se tornando o padrão por conta da quantidade de dados que as organizações precisam gerenciar.

Big data significa coisas diferentes para diferentes empresas. Por exemplo, isso pode significar que empresas queiram analisar todos os textos (e.g. contratos e mensagens de mídia social), mensagens de voz (e.g. conversas entre controladores aéreos e pilotos e operadores de call center), imagens (e.g. danos ao carro devido a acidentes), e vídeos (e.g. de câmeras de segurança e câmeras em aeroportos e lojas). *Big data* também pode significar que mais processos de negócios são detalhadamente monitorados, resultando em grandes quantidades de dados minuciosos, indicando o progresso e o sucesso desses processos. Novamente, isso impactará a tecnologia e as ferramentas a serem usadas. *Big data* também pode se referir a dados de *streaming* que talvez precisem ser analisados logo depois de serem gerados.

Sem dúvida, o que quer que *big data* signifique para uma organização, a nova e crescente quantidade de dados armazenados, processados e analisados terá um grande impacto na tecnologia usada na nova arquitetura escolhida.

O big data tem um grande impacto sobre a tecnologia utilizada.

Suporte bi-modal – A maioria dos sistemas analíticos existentes é desenvolvida para suportar formas tradicionais de relatórios, interfaces e análises. Essas formas de análise devem ser controláveis, auditáveis e formalmente testadas. Ultimamente, mais e mais departamentos de empresas desejam desenvolver seus próprios relatórios usando *ferramentas de análise com auto-atendimento*. Esses relatórios feitos pelos próprios usuários, ou no auto-atendimento, podem ter caráter exploratório e ter uma curta vida útil. Alguns relatórios são desenvolvidos para uso único, para resolver apenas um problema.

O termo *bi-modal* refere-se à implementação dessas soluções experimentais mais ágeis em combinação com as formas mais estáveis, tradicionais e formais de geração de relatórios. Essa combinação de relatórios de auto-atendimento e relatórios formais normalmente leva a um choque de culturas e tecnologias. É um desafio desenvolver uma arquitetura de dados única que suporte o formato bi-modal satisfatoriamente. É como projetar um carro que deve ser muito rápido, mas que também seja capaz de transportar uma carga pesada. Esses dois aspectos parecem ser opostos.

A implementação bimodal em uma nova arquitetura de dados é um desafio.

Independência e Interoperabilidade da Plataforma em Nuvem – Levar aplicativos e sistemas para a nuvem já é uma coisa comum. Existem muitas plataformas em nuvem, incluindo as da Amazon, Google e Microsoft. Todas têm seus pontos fortes e fracos. Uma nova arquitetura de dados deve ser o mais independente possível da plataforma em nuvem, a fim de permitir que a empresa tenha a flexibilidade de alternar entre as tecnologias que melhor atendem aos seus requisitos.

A independência e a interoperabilidade de uma plataforma em nuvem são requisitos fundamentais.

Além disso, empresas podem usar várias plataformas em nuvem. Nesse caso, a nova arquitetura de dados deve suportar a *interoperabilidade da plataforma em nuvem*. Por exemplo, ele deve ser transparente para relatórios que exigem dados de dois sistemas em execução em duas plataformas em nuvem diferentes. A independência e a interoperabilidade da plataforma em nuvem são requisitos essenciais para arquiteturas de dados modernas.

Ciência de Dados – A ciência de dados permite que as empresas criem modelos analíticos para apoiá-las com decisões melhores e / ou mais rápidas. Técnicas como estatística, *deep learning*, *machine learning* e IA são implementadas para criar esses modelos analíticos. Primeiro: uma nova arquitetura de dados deve ajudar os cientistas de dados a desenvolver facilmente novos modelos de forma experimental e investigativa. Segundo: os estudos mostram que os cientistas de dados gastam 80% do tempo na preparação de dados e apenas 20% no trabalho analítico real. Uma nova arquitetura de dados deve ajudar a reduzir o tempo necessário para adquirir e preparar dados.

Novas arquiteturas devem reduzir o tempo de preparação dos dados.

Streaming de Dados – A necessidade de dados em tempo real pode fazer com que seja necessário processar dados de *streaming*, para lidar, por exemplo, com dados da IoT. Os dados são passados ou *empurrados* diretamente da fonte para alimentar aplicativos de *streaming* em tempo real, que transformam ou reagem a novas informações. Novas arquiteturas precisam acomodar dados de streaming com sistemas que possam suportar cenários assíncronos de publicação e assinatura semelhantes a uma fila de mensagens ou sistema de mensagens corporativo. Pode ser que isso tenha que acontecer sem qualquer forma de armazenamento temporário de dados. Muitas perguntas precisam ser respondidas, como por exemplo como será feita a limpeza instantânea de dados, com que rapidez os modelos analíticos podem ser implantados, se a infraestrutura consegue processar a carga de trabalho dos dados em *streaming* e se esses dados também são armazenados, e ainda se a tecnologia de banco de dados consegue acompanhar a taxa de inclusão de dados produzida por esses fluxos, e assim por diante. Uma nova arquitetura de dados precisa oferecer suporte ao *streaming* de dados.

Novas arquiteturas de dados devem suportar streaming de dados.

Interfaces Orientadas por Serviços – Nem todos os consumidores de dados são humanos. Cada vez mais, os dados precisam ser disponibilizados por meio de interfaces de serviço independentes e sem bandeira. Por exemplo, os serviços devem ser desenvolvidos para atender o público externo, permitindo que outras partes, como fornecedores e representantes, criem aplicativos que acessem esses dados. Nesse caso, o consumidor de dados não é humano, mas sim um aplicativo. Da mesma forma, as plataformas nas quais a nova arquitetura de dados irá operar também precisam ser orientadas a serviços, implantando micro serviços e repositórios.

Dados precisam estar disponíveis nas de interfaces de uso.

Uma consequência desse tipo de uso de dados é que ele pode levar a uma carga de trabalho de *query* enorme e altamente imprevisível. As tecnologias usadas na nova arquitetura de dados devem ser poderosas e escalonáveis o suficiente para dar o suporte necessário.

Esses são alguns dos desafios que as empresas enfrentam ao projetar suas novas arquiteturas de dados e ao selecionar novas tecnologias, ferramentas e técnicas.

3 A Virtualização de Dados como Catalisador para Arquiteturas de Dados Modernas

Os sistemas de entrega de dados existentes podem ser modernizados e ficar *à prova de futuro* de muitas maneiras diferentes e com diversas tecnologias e ferramentas. Uma tecnologia que pode ajudar com uma modernização quase perfeita de uma arquitetura de dados é *virtualização de dados*. Uma definição comum, um tanto descritiva¹ de virtualização de dados é: *A virtualização de dados oferece aos consumidores uma solução unificada, abstraída e visão encapsulada para query e manipulação dados armazenados em um conjunto heterogêneo de data stores.*

O risco comum da modernização é que possivelmente os sistemas de relatórios existentes fiquem temporariamente off-line. A virtualização de dados permite que as arquiteturas sejam modernizadas sem perturbar a carga de trabalho de análise e relatório existente. Esta seção lista alguns dos recursos da virtualização de dados que podem facilitar a modernização da arquitetura.

Desempenho Aprimorado por Meio de Query Pushdown – Quando servidores de virtualização de dados recebem uma solicitação, eles podem determinar quanto do processamento é executado pelo próprio servidor de virtualização de dados e quanto pelas fontes de dados subjacentes. O processo de “empurrar” a consulta para a fonte de dados é chamado *query pushdown*; Veja a figura 1 A vantagem dessa abordagem é que todo o potencial de consulta dessa plataforma de banco de dados é usado. Isso é especialmente importante para novas plataformas, como o Hadoop e alguns dos sistemas analíticos SQL, porque oferecem recursos rápidos de processamento de consultas paralelas. A função de *query pushdown* pode acelerar o desempenho, explorando o poder da consulta.

Através do query tipo pushdown todo o potencial de um banco de dados é usado.

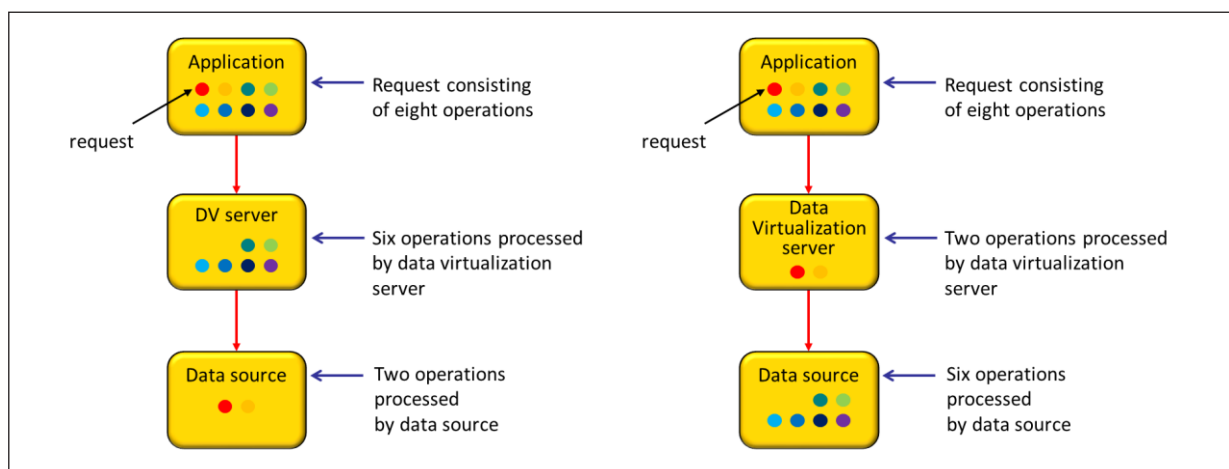


figura 1 Usando query pushdown para explorar todo o potencial de consulta na fonte de dados subjacente.

Segurança de Dados Aprimorada pela Centralização – Quando os dados são dispersos em várias plataformas de banco de dados diferentes, os responsáveis pela segurança precisam lidar com diferentes sistemas. Isso significa que, para cada usuário, diferentes tipos de especificações de autorização, autenticação e anonimato precisam ser definidos para indicar quais usuários tem permissão para usar os dados. Potencialmente, esse é um cenário de pesadelo para a segurança. Como todas essas especificações são mantidas em sincronia? Tome como exemplo um funcionário que pediu demissão. Ele deve ser removido da lista de usuários de todos os sistemas aos quais ele teve acesso.

A virtualização de dados dá suporte a uma camada de segurança

¹ RF van der Lans, *Virtualização de dados: Escritos Seleccionados*, Lulu.com, setembro de 2019; consulte <http://www.r20.nl/DataVirtualizationBook.htm>

Com a virtualização de dados, uma camada centralizada de especificação de segurança pode ser definida e aplicada em todas as fontes de dados. A capacidade de especificar quem tem permissão para ver qual parte dos dados precisa ser definida apenas uma vez.

Anonimização de Dados Aprimorada com Filtros — Novos regulamentos e leis estão forçando as empresas a implementar regras de anonimato em relação à *informação pessoalmente identificável* (PII). Algumas PII precisam ser mascaradas, outras PII devem ser ilegíveis de forma que não possam ser rastreadas até uma pessoa específica. Os servidores de virtualização de dados dão o suporte aos filtros e recursos de transformação para implementar muitas dessas regras de anonimização. Isso pode ser feito sem afetar os bancos de dados originais. Assim como as regras de segurança de dados, as regras de anonimização são definidas centralmente e podem operar com dados virtualizados de qualquer tipo de plataforma de banco de dados.

A virtualização de dados suporta recursos para anonimização de dados.

Observe que os dados armazenados permanecem inalterados, sem máscara e sem distorção. Evidentemente, se houver a necessidade de fazer uma consulta aos dados diretamente e ignorar o servidor de virtualização, as regras de anonimização deverão ser implementadas na própria plataforma de banco de dados.

Melhor Velocidade de Desenvolvimento com Especificações Centralizadas — Para transformar dados provenientes de uma variedade de fontes em dados significativos para usuários, eles precisam ser processados, o que significa que precisam ser filtrados, limpos, integrados, agregados, calculados e assim por diante. Nos sistemas analíticos tradicionais, essas *especificações de transformação* são implementadas em todo o sistema, nos relatórios, nos programas ETL, nos próprios bancos de dados e como fragmentos de código nos programas manuscritos. Isso é prejudicial à produtividade e dificulta a manutenção.

Os servidores de virtualização de dados permitem que a maioria das especificações de transformação seja definida centralmente em um *repositório*; Veja a figura 2) As especificações são definidas uma vez e podem ser reutilizadas, melhorando a velocidade de desenvolvimento e facilitando a manutenção. Além disso, os recursos internos de análise de origem e impacto permitem que os desenvolvedores (e usuários) estudem as especificações e vejam como elas estão relacionadas.

Especificações de transformação podem se definir centralmente com um servidor de virtualização de dados.

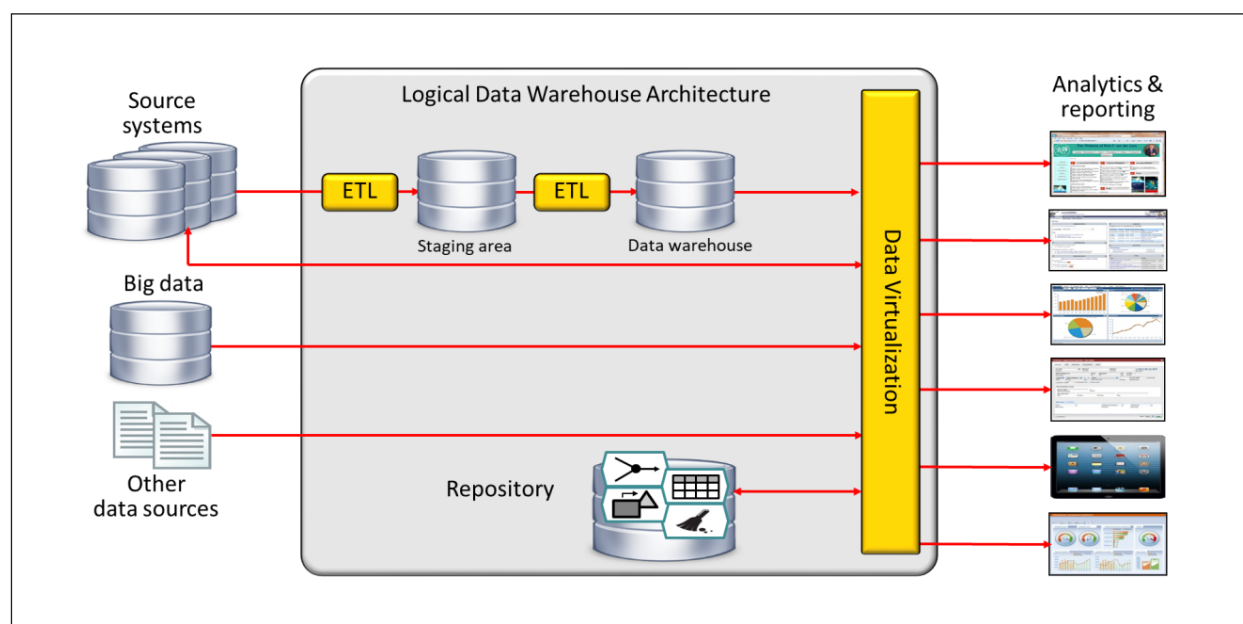


Figura 2 Os servidores de virtualização de dados armazenam todas as especificações de transformação em um repositório central.

Migração Fácil de Banco de Dados com Independência – Os servidores de virtualização de dados podem extrair dados de todos os tipos de servidores de banco de dados, incluindo a maioria dos produtos SQL, Hadoop, planilhas, fontes JSON e XML e aplicativos. As ferramentas de geração de relatório podem usar um servidor de virtualização de dados para acessar qualquer um desses servidores de banco de dados usando a mesma linguagem. Seja acessando dados armazenados no Hadoop, Oracle ou SnowflakeDB, todos esses sistemas podem ser consultados usando as mesmas instruções. Em outras palavras, a virtualização de dados torna a tecnologia do banco de dados de relatórios independente.

Esse recurso pode ser usado quando houver necessidade de migrar dados para uma plataforma moderna; Veja a figura 3) Na fase 1, o servidor de virtualização de dados é colocado entre os relatórios e o banco de dados existente. Na fase 2, os dados são copiados para o novo banco de dados e na fase 3 o servidor de virtualização de dados é redirecionado para acessar o novo banco de dados. Tudo isso pode ser feito sem fazer nenhuma ou pequenas alterações nos relatórios. Veja que a migração de dados pode ser feita tabela a tabela ou todos os dados podem ser migrados em uma grande etapa.

A independência do banco de dados é um recurso valioso para se mover gradual e continuamente para uma nova arquitetura de dados.

A virtualização de dados torna a tecnologia do banco de dados de relatórios independente e, portanto, mais fácil de

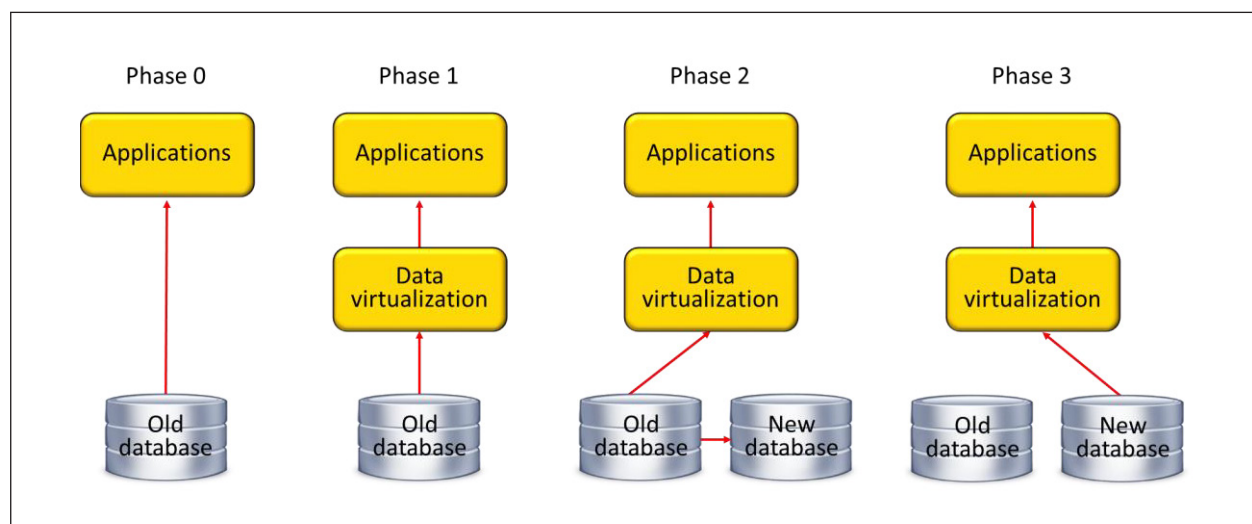


Figura 3 Migrando para uma plataforma moderna.

Dissociação com Abstração de Dados – A virtualização de dados pode ser usada para fornecer uma *camada virtual de abstração dados* que é colocada acima de toda a arquitetura de dados, dissociando os relatórios das tecnologias usadas para armazenar e processar os dados; veja a figura 4. Com a virtualização de dados atuando como uma camada de abstração, podem ser feitas alterações na camada de armazenamento de dados sem afetar os relatórios. Por exemplo, diferentes tecnologias podem ser usadas para copiar os dados de um banco para outro, as camadas do banco de dados podem ser ignoradas para acelerar a entrega dos relatórios, então diferentes técnicas usadas para projetar aquele de banco de dados podem ser usadas para rastrear o histórico de dados e outras soluções de *data warehouse* podem ser vinculadas a essa arquitetura, mas ainda serão vistas como uma arquitetura integrada, e assim por diante. Em outras palavras, com a virtualização de dados, a camada de armazenamento pode ser modernizada sem afetar a geração de relatórios.

Observe que essa abordagem de empacotamento também pode ajudar a integrar novos sistemas de relatórios provenientes da aquisição de outra empresa.

A virtualização de dados dissocia relatórios da arquitetura de dados existente.

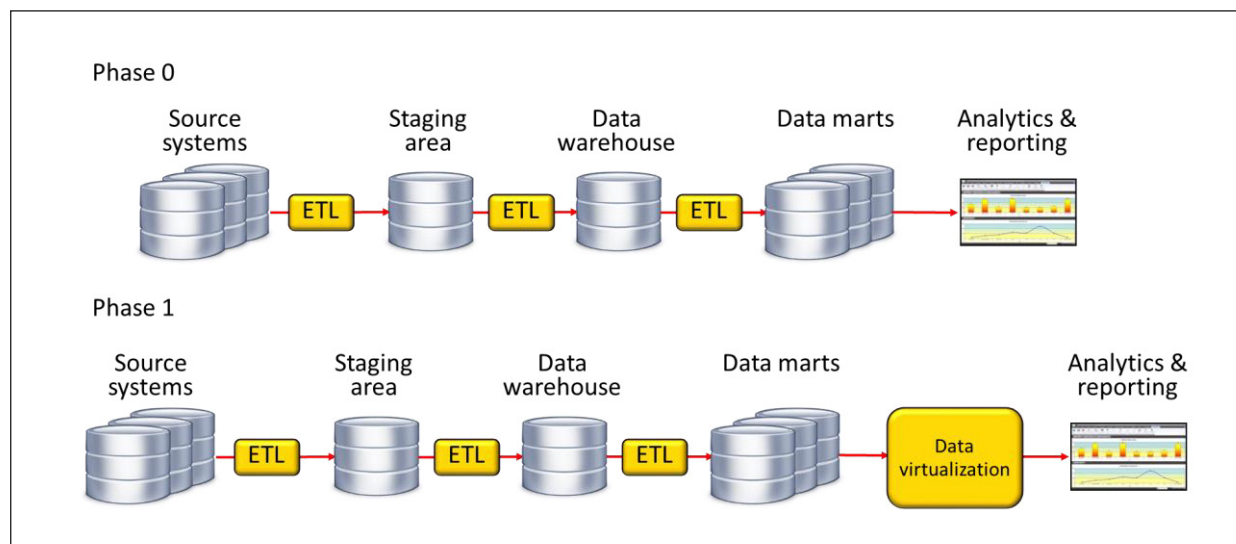


Figura 4 Empacotando uma arquitetura com virtualização de dados.

Maior Transparência com um Catálogo de Dados – Cada vez mais, as organizações estão recorrendo aos catálogos de dados como um meio para armazenar ativos de dados valiosos para recuperação rápida. Na virtualização de dados, um catálogo pode ser usado para armazenar os meta-dados ou visualizações virtuais que foram definidas, descritas e documentadas. Além disso, eles podem ser marcados e enriquecidos com descrições adicionais para torná-los mais compreensíveis e fáceis de consultar. Desenvolvedores, usuários corporativos e cientistas de dados têm benefícios práticos de um catálogo de virtualização de dados, pois podem usá-lo para procurar elementos de dados necessários para seus relatórios ou algoritmos, e é especialmente útil para ambientes de investigação ad-hoc.

Os servidores de virtualização de dados suportam um catálogo centralizado e que pode ser usado para consultas.

Adoção Fácil na Nuvem Através da Abstração – A modernização das arquiteturas de dados pode envolver a adoção de uma *plataforma em nuvem*, como as da Amazon, Google ou Microsoft. Tecnicamente, isso pode significar muitas coisas. Primeiro, isso pode significar que um banco de dados é migrado para a nuvem. Essa migração pode ficar oculta para todos os relatórios e aplicativos por meio de um servidor de virtualização de dados. Nesse caso, o servidor de relatórios e de virtualização de dados ainda roda fisicamente no local e esse último cuida de todo o acesso ao banco de dados baseado em nuvem; (veja a figura 5) Essa migração dos dados não precisa ser executada em uma única etapa: pode ser feita gradualmente. O servidor de virtualização de dados lida com as consultas que ingressam nos dados armazenados na nova plataforma baseada em nuvem e nos dados que ainda permanecem no banco de dados local. Especialmente se os bancos de dados forem grandes, é preferível uma migração gradual ou em fases para garantir 100% de disponibilidade.

Uma migração contínua e sem interrupções para plataformas em nuvem com a virtualização de dados.

Segundo, se uma empresa quer usar a tecnologia analítica de acesso rápido ao banco de dados em nuvem, um servidor de virtualização de dados pode ocultar o fato de que os dados não são só estão em nuvem, mas também estão em outra plataforma de banco de dados. O servidor de virtualização de dados oculta o fato de que os dados são acessíveis por meio de uma linguagem SQL ligeiramente diferente e usa o *query pushdown* para explorar o potencial da nova plataforma de banco de dados.

Terceiro, a empresa começou a usar aplicativos de nuvem proprietários, como Salesforce.com e NetSuite. Os dados produzidos nesses aplicativos em nuvem devem estar acessíveis para geração de relatórios e integração com outros bancos de dados. Os servidores de virtualização de dados acessam esse aplicativos em nuvem como se fossem bancos de dados.

Quarto, se toda a arquitetura precisar ser migrada para a nuvem, os recursos da camada de abstração de dados virtuais o permitirão fazer isso, passo a passo.

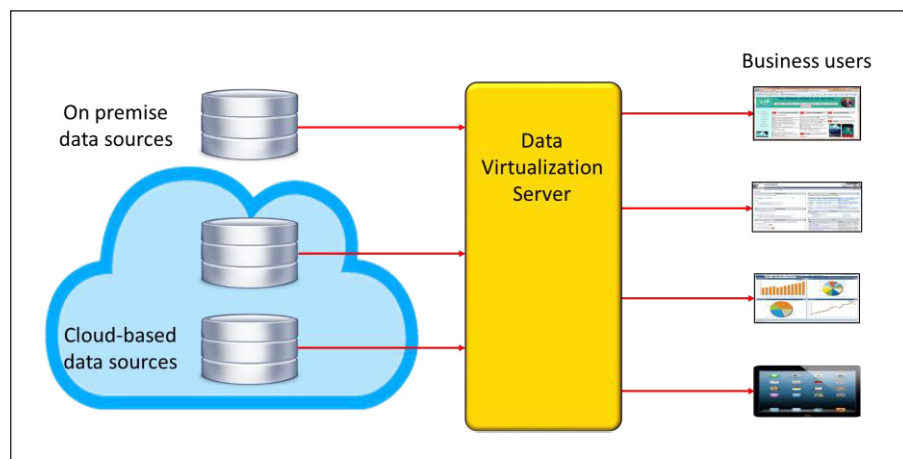


Figura 5 Como ocultar a migração de um banco de dados para uma plataforma em nuvem com a virtualização de dados.

Entrega de Dados em Tempo Real com o Acesso sob Demanda – Servidores de virtualização de dados podem acessar qualquer tipo de banco de dados, incluindo os bancos de dados pertencentes a sistemas transacionais. Servidores de virtualização de dados podem ajudar relatórios a mostrar dados em tempo real armazenados nesses bancos de dados transacionais; Veja a figura 6 O desafio de desenvolver esses relatórios é não permitir que interfiram nas transações em execução, e nem atrasar a geração de relatórios por conta disso. Os servidores de virtualização de dados oferecem suporte a recursos que resolvem isso, como a otimização avançada de query, a capacidade de forçar o servidor de virtualização de dados a executar as consultas precisas que se deseja e armazenar em cache os dados transacionais.

A virtualização de dados permite entregar dados em tempo real para relatórios ao acessar sistemas transacionais

Além disso, os recursos de query pushdown permitem que um servidor de virtualização de dados use toda a potencial da plataforma de banco de dados subjacente. E se o usuário acessar grandes quantidades de dados, a virtualização de dados usará os recursos de processamento de query paralelas para acelerar o desempenho. O armazenamento em cache também pode ajudar a minimizar a interferência, ao guardar alguns dados que quase nunca são alterados, como tabelas de referência.

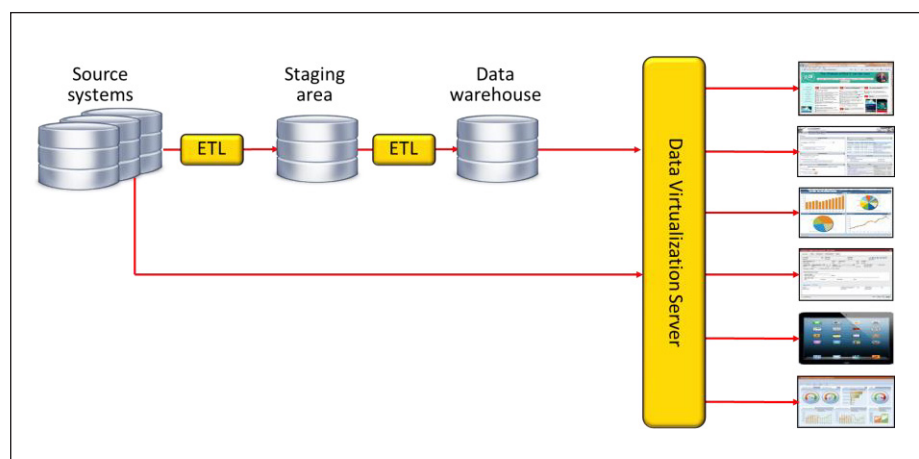


Figura 6 A virtualização de dados pode apresentar dados em tempo real, acessando diretamente os sistemas de transações.

Em resumo, a virtualização de dados ajuda a modernizar arquiteturas de dados existentes de maneira uniforme e gradual. Não há necessidade de uma substituição completa de toda a arquitetura, o que aumentaria os riscos de uma mudança.

4 Outras Soluções para Modernização da Arquitetura de Dados

Outros conceitos e tecnologias são frequentemente propostos para ajudar a modernizar uma arquitetura de dados. Esta seção descreve os inconvenientes e benefícios para cada um deles. Além disso, esta seção explica como a virtualização de dados pode ajudar a simplificar a adoção de tais conceitos em uma arquitetura de dados nova e à prova de futuro.

Data Lake — Uma forma de modernização é copiar todos (ou a maioria) dos dados originais para um *data lake* central. Nesse caso, o *data lake* se torna o ponto focal de todas as formas de acesso e entrega de dados. Isso pode ser feito basicamente de duas maneiras. Primeiro, algumas organizações incluem o *data lake* entre os sistemas de origem e o *data warehouse*; ver Figura 7. Dessa maneira, o *data lake* age como o que antes era chamado, um *data store* operacional. Esse *data lake* é usado por usuários comerciais, como cientistas de dados, diretamente e pelos processos de ETL que copiam os dados do *data lake* para o *data warehouse*. A tecnologia de reprodução ou *streaming* é usada para manter o *data lake* atualizado. De ponto de vista do histórico completo e integração de dados, essa é uma alternativa interessante. No entanto, seu ponto fraco é não conseguir disponibilizar dados em tempo real.

O data lake é o ponto focal de todas as formas de uso de dados.

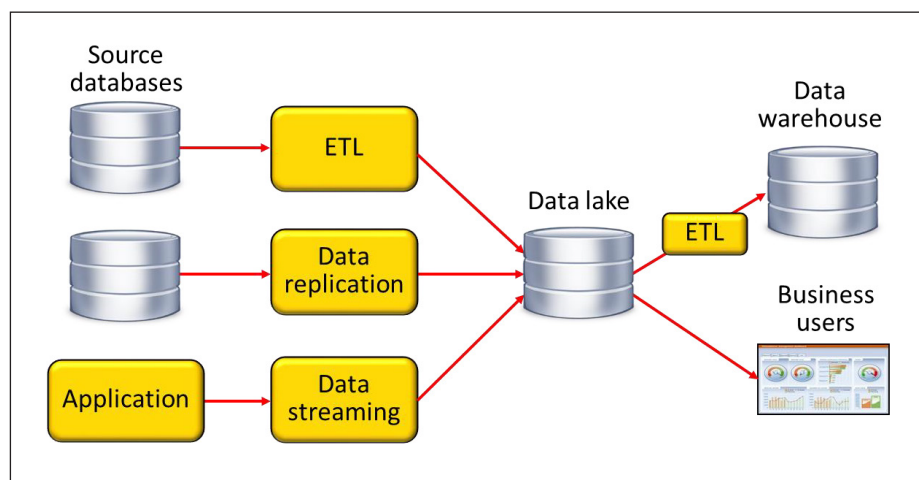


Figura 7 Inclusão de um *data lake* em uma arquitetura de dados existente.

Segundo, o *data lake* é introduzido como um repositório central para armazenar todos os dados da empresa e dados externos. Aqui, o *data lake* substitui o *data warehouse* existente ou sistemas similares. A implementação de um *data lake* tem um enorme impacto em boa parte da arquitetura de dados. Os relatórios podem ter que ser alterados, assim como os processos ETL.

Para fazer isso funcionar, todos os dados precisam ser movidos para o *data lake*. A solução seria vincular o *data lake* à arquitetura de dados existente com a virtualização de dados. A introdução de um *data lake* em uma arquitetura de dados existente pode levar à migração de relatórios que rodam atualmente em um *data warehouse*, um *data mart* ou outro banco de dados. Pode não ser uma migração simples e direta. Os dados em um *data lake* geralmente são estruturados de maneira diferente do que em um dos outros bancos de dados mencionados. Isso pode significar que o código que acessa os dados deve ser reconstruído. Com um servidor de virtualização de dados, essa estrutura de dados alterada fica oculta para os relatórios. Isso simplifica a introdução do *data lake* em uma arquitetura de dados existente.

Plataformas na Nuvem — As plataformas em nuvem provaram seu valor para as empresas. Evidentemente, diferentes organizações obtêm benefícios comerciais distintos. Um valor comercial dominante é que as plataformas em nuvem trazem é *desafogar* a carga de trabalho. As empresas não precisam mais gerenciar e monitorar seus ambientes de *hardware*. Não há razão para equipar salas especiais de um *data center* com *hardware* e pisos elevados, lidar com

As plataformas em nuvem desafogam as organizações e oferecem escalabilidade ilimitada.

linhas de comunicação rápidas, controle a temperatura e cumprimento de acordos de nível de serviço adequados. Um segundo benefício comercial é a disponibilidade quase ilimitada de capacidade de processamento e armazenamento. Isso oferece à empresa escalabilidade fácil e ilimitada, poder computacional ou de armazenamento adicional, que pode estar disponível em segundos, quando necessário.

As plataformas em nuvem desempenham um papel importante na modernização das arquiteturas de dados. Algumas organizações levaram seus componentes de software existentes para a nuvem. Outras tentam explorar as plataformas na nuvem de maneira ideal adotando também *tecnologias nativas* desenvolvidas e otimizadas especificamente para plataformas em nuvem, como Google BigQuery e Amazon Athena. Se os relatórios são desenvolvidos diretamente nesses sistemas, um certo *lock-in* acontece. A dependência de uma plataforma de nuvem específica aumenta quando essas tecnologias nativas de nuvem são implantadas. Em outras palavras, essas tecnologias dificultam novamente a migração futura dos clientes. Conforme indicado neste *whitepaper*, a dependência dessas tecnologias otimizadas, porém proprietárias, pode ser minimizada usando a virtualização de dados, mantendo todos os benefícios de desempenho, escalabilidade e flexibilidade.

Muitas empresas já chegaram ao estágio em que usam várias plataformas em nuvem. Por exemplo, rodam o Salesforce.com em uma nuvem, alguns de seus dados são armazenados no Hadoop na AWS e instalaram o SAP no Azure. Independentemente do restante da arquitetura de dados, é necessária uma solução para integrar os dados provenientes dessas múltiplas plataformas em nuvem. Em termos de tecnologia, este é um *query de nuvem híbrida*. A virtualização de dados pode lidar com esse tipo de consulta e sem precisar mostrar que os dados são armazenados em diferentes plataformas em nuvem.

Análise de Dados em Auto-atendimento — Às vezes, os novos requisitos de negócios feitos pelos usuários não são implementados ao modernizar a arquitetura, mas sim ao dar a eles os recursos para o desenvolvimento de relatórios. A auto-atendimento para análise de dados é um bom exemplo disso. Antigamente, os relatórios antigos eram sempre desenvolvidos por especialistas em TI. Atualmente, alguns dos produtos são suficientemente intuitivos para os usuários corporativos desenvolverem os próprios relatórios. Em outras palavras, se modernização significa que recursos e o formato de apresentação ou relatórios e análises precisam ser aprimorados, isso é definitivamente um caminho interessante.

*Análise em auto-atendimento
diz respeito a dar
recursos de geração
de relatórios
ao usuários.*

De fato, algumas dessas ferramentas agora oferecem suporte à funcionalidade de integração. Usuários podem agora facilmente definir como integrar dados de várias fontes. Isso às vezes é chamado de *citizen integration*. Termos como *disputa* e *preparação de dados* referem-se a funcionalidades semelhantes. Todos estes termos significam que mais recursos de integração e preparação estão disponíveis para os usuários. Observe que o fator limitante é que eles ainda devem acessar bancos de dados desenvolvidos e pertencentes à TI. Então, eles ainda estão limitados por esse mundo.

A análise de dados por auto-atendimento tem muito a oferecer a uma empresa. A grande desvantagem é que muitas especificações para integração, agregação, limpeza e filtragem são desenvolvidas de maneira isolada. Isso possivelmente leva a resultados de relatórios inconsistentes e incorretos, diminui a produtividade dos usuários, prejudica a manutenção e a qualidade das especificações desenvolvidas por eles é mais difícil de testar e raramente pode ser auditada.

A virtualização de dados pode minimizar esse risco, permitindo que muitas dessas especificações sejam definidas centralmente. Nesse caso, as especificações podem ser compartilhadas e reutilizadas por todos os relatórios feitos em auto-atendimento. Eles podem até ser compartilhados entre diferentes ferramentas. Isso resolve os problemas descritos.

Aproveitando os Meta Dados — Em muitos projetos de modernização, o ponto focal é a melhoria da manutenção e entrega de meta-dados. Existe em toda a organização uma necessidade crescente de transparência, governança e meta-dados que possam ser consultados. O foco em tais projetos não é apenas o desenvolvimento de uma arquitetura de dados ‘melhor’, mas também a capacidade de descrever e pesquisar meta-dados, especialmente meta-dados comerciais.

Existem algumas soluções alternativas. Muitas dessas soluções permitem criar um catálogo de dados independente no qual todos os meta-dados são armazenados, definidos, documentados e vinculados. Usuários e desenvolvedores podem navegar neste catálogo, pesquisar termos específicos e ver de onde vieram. Infelizmente, com essas soluções independentes, a maioria dos meta-dados nem sempre é atualizada automaticamente quando partes dos sistemas são excluídos ou alterados. Para atualizá-los, uma ação manual ou parcialmente automatizada

deve ocorrer. O risco é que isso seja omitido e o catálogo de dados lentamente fique desatualizado e eventualmente obsoleto.

Servidores de virtualização de dados, como a Plataforma Denodo, suportam um catálogo de dados integrado. Quando um objeto de dados é definido, ele pode ser descrito com meta-dados técnicos e de negócios. As especificações de origem são mantidas atualizadas automaticamente. A exclusão de um elemento de dados altera automaticamente o catálogo. Em outras palavras, o catálogo é mais integrado aos sistemas operacionais do que as soluções independentes e, portanto, sempre atualizado.

A Denodo Platform é uma das poucas plataformas de virtualização de dados na qual elementos de meta-dados podem ser vinculados aos relatórios. Evidentemente, este catálogo de dados não conterá todos os meta-dados.

O Data Warehouse Lógico – A abordagem mais óbvia da modernização é aquela em que a arquitetura atual é ampliada ou alguns componentes são substituídos. Por exemplo, com o *data warehouse lógico* a arquitetura clássica existente pode ser agrupada; Veja a figura 8) Nesta nova arquitetura, todos os consumidores acessam os dados por meio do servidor de virtualização de dados. Alguns dos dados acessados podem vir do *data warehouse* existente, outros podem vir de novas fontes de *big data* ou de qualquer outra fonte de dados. O servidor de virtualização de dados não precisa mostrar a origem dos dados. Isso permite uma introdução gradual da virtualização de dados e uma migração gradual para uma arquitetura completamente nova. Por exemplo, quando todos os consumidores acessam todos os dados por meio da camada de virtualização de dados, fica mais fácil fazer alterações no *data warehouse*. A estrutura de dados ou a maneira como *data warehouse* de dados é carregado pode ser alterada sem afetar os relatórios.

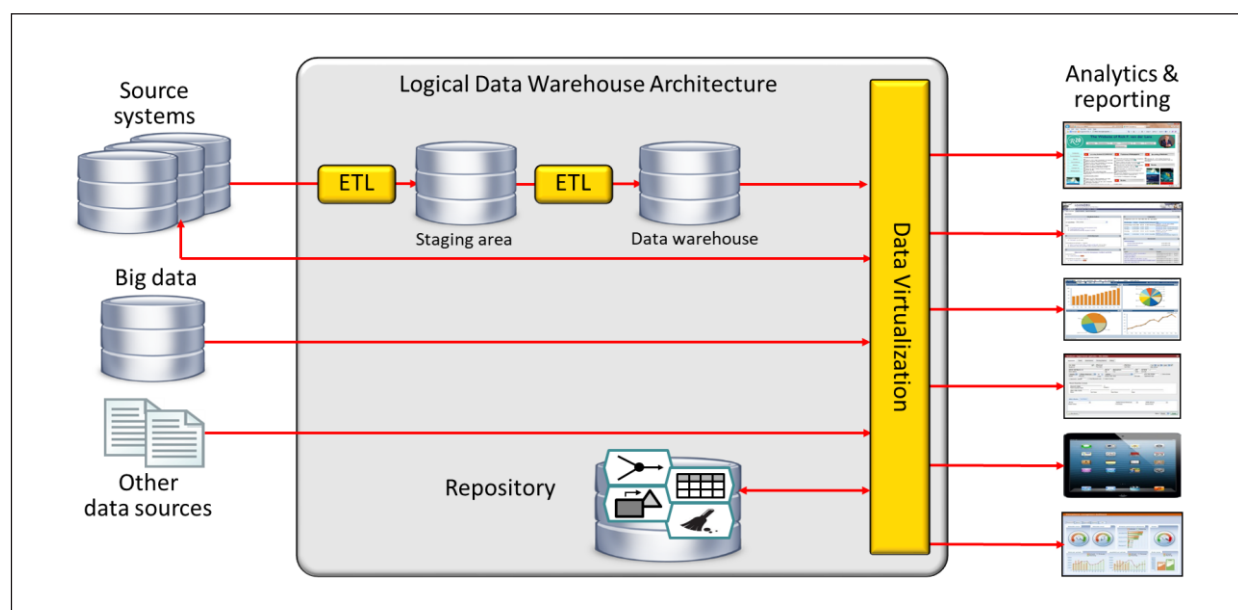


Figura 8 Visão geral da arquitetura lógica do data warehouse.

5 Comentários finais

É ingênuo pensar que uma arquitetura de dados existente pode ser desligada um dia e a nova ligada no dia seguinte. Por inúmeras razões, uma abordagem do tipo desfazer e substituir tudo não é realista na maioria das situações. Por exemplo, se a arquitetura suporta acesso a dados 24 horas por dia, 7 dias por semana, quando a troca deve ser feita? Ou, a mudança pode envolver fechar o sistema antigo e copiar os dados para o novo sistema antes que eles possam se tornar operacionais e, especialmente se se tratar de *big data*, apenas o processo de cópia em si pode levar muito tempo.

Uma abordagem de desfazer tudo e substituir toda a arquitetura de dados não é uma boa opção.

A migração ideal é contínua e gradual, quase imperceptível pelos consumidores de dados existentes. A migração ideal evolui suavemente da arquitetura de dados antiga para a nova. Passo a passo, pequenos módulos das arquiteturas de dados existentes são substituídos por novos. Graças à migração gradual, a modernização rápida permite à empresa competir na era digital atual e realizar seu sonho de transformação digital.

Os muitos recursos de agrupamento e abstração dos servidores de virtualização de dados oferecem uma migração gradual e contínua. Novos módulos podem ser facilmente adicionados à arquitetura existente, e os antigos podem ser substituídos. Além disso, a adoção de plataformas em nuvem se torna mais fácil e é menos suscetível a riscos quando a virtualização de dados é capaz de ocultá-los. Também torna a nova arquitetura de dados menos dependente de uma plataforma em nuvem e de sua tecnologia nativa.

Outras tecnologias e abordagens usadas para modernização, como *data lakes*, auto-atendimento de BI, plataformas em nuvem e catálogos de dados, também podem ser relevantes. Nesse caso, a virtualização de dados pode funcionar com todos eles. Na verdade, na maioria das vezes a virtualização de dados pode amplificar seus benefícios.

A combinação da virtualização de dados com data lakes, auto-atendimento de BI, plataformas em nuvem e catálogos de dados tem um

Sobre o autor

Rick van der Lans é um analista de sistemas independente, consultor, autor e conferencista internacionalmente respeitado, especializado em data warehousing, business intelligence, big data, tecnologia de banco de dados e virtualização de dados. Ele trabalha na R20 / Consultancy (www.r20.nl), que fundou em 1987. Em 2018, ele foi selecionado como o sexto analista de sistemas de BI mais influente do mundo pela onalytica.com².

Ele apresentou inúmeros seminários, webinários e palestras em conferências renomadas no setor. Por muitos anos, atuou como presidente do conselho da conferência anual European Enterprise Data and Business Intelligence Conference, em Londres e da Data Warehousing and Business Intelligence Summit.

Rick ajuda clientes em todo o mundo a criar suas arquiteturas e soluções de data warehouse, big data e business intelligence e auxilia na seleção dos produtos certos. Ele foi influente na introdução da nova arquitetura lógica de data warehouse em todo o mundo, ajudando organizações a desenvolver sistemas de inteligência de negócios mais ágeis. Ele apresentou a arquitetura de Business Intelligence, chamada Data Delivery Platform, em 2009, em vários artigos todos publicados na B-eye-Network.com.

Ele é autor de vários livros sobre computação, incluindo o novo *Data Virtualization: Selected Writings*⁴ e *Data Virtualization for Business Intelligence*⁵. Alguns desses livros estão disponíveis em outros idiomas. Livros como o popular *Introduction to SQL* estão disponíveis em inglês, holandês, italiano, chinês e alemão e são vendidos em todo o mundo. Ao longo dos anos, ele escreveu centenas de artigos e blogs para jornais e sites e foi autor de muitos *white papers* educacionais e populares para uma longa lista de fornecedores. Ele foi o autor do primeiro livro disponível sobre SQL6, titulado *Introduction to SQL*, que foi traduzido para vários idiomas com mais de 100.000 cópias vendidas.

Para mais informações, visite www.r20.nl ou envie um e-mail para rick@r20.nl. Você também pode entrar em contato com ele via LinkedIn e Twitter (@Rick_vanderlans).

Embaixador da Axians Business Analytics Laren (anteriormente Kadenza): Essa empresa de consultoria é especializada em business intelligence, gerenciamento de dados, *big data*, *data warehousing*, virtualização e análise de dados. Nesta função em meio período, Rick trabalha em conjunto com os consultores em muitos projetos. Suas experiências e ideias conjuntas são compartilhadas em seminários, webinários, blogs e *whitepapers*.

Sobre a Denodo

A Denodo é líder em virtualização de dados, fornecendo integração de dados ágil e de alto desempenho, abstração e serviços de dados em tempo real, na mais ampla gama de fontes corporativas, em nuvem, em big data e em dados não estruturados pela metade do custo das abordagens tradicionais. Os clientes da Denodo em todos os principais setores obtiveram retorno sobre seu investimento e agilidade comercial significativos, permitindo acesso mais rápido e fácil a informações comerciais unificadas para BI ágil, análise de *big data*, integração de Web e nuvem, aplicativos de visualização única e serviços de dados corporativos. A Denodo tem financiamento saudável, é lucrativa e de capital fechado. Para mais informações, visite www.denodo.com.

² Onalytica.com, *Business Intelligence - Top Influencers, Brands and Publications* Junho de 2018; consulte <http://www.onalytica.com/blog/posts/business-intelligence-top-influencers-brands-publications/>

³ Consulte <http://www.b-eye-network.com/channels/5087/view/12495>

⁴ RF van der Lans, *Data Virtualization: Selected Writings*, Lulu.com, setembro de 2019; consulte <http://www.r20.nl/DataVirtualizationBook.htm>

⁵ RF van der Lans, *Data Virtualization for Business Intelligence Systems*, Morgan Kaufmann Publishers, 2012.

⁶ RF van der Lans, *Introduction to SQL; Mastering the Relational Database Language*, quarta edição, Addison-Wesley, 2007.