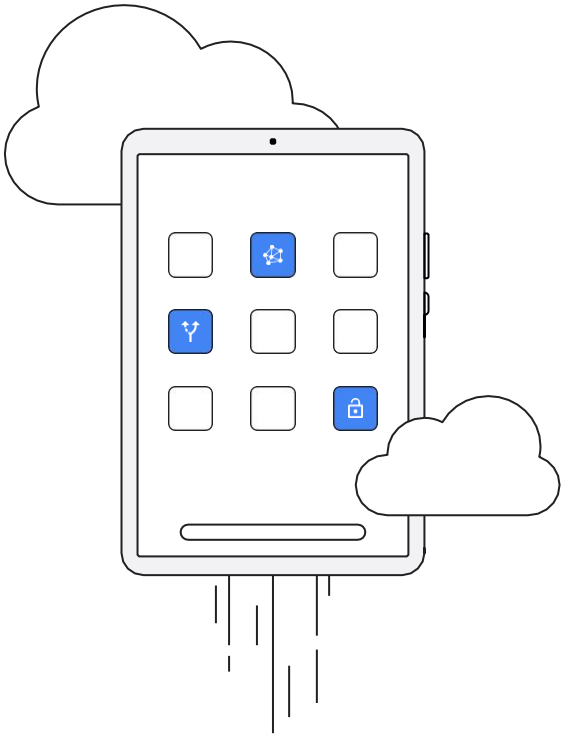
Google Cloud

O futuro dos dados será unificado, flexível e acessível

Empresas e startups de tecnologia estão descobrindo que para ter sucesso:

- Os dados de toda a empresa, incluindo de fornecedores e parceiros, precisam ser *unificados*. Para isso, é preciso desbloquear os dados não estruturados e eliminar os silos de organizações e tecnologias.
- O stack de tecnologia precisa ser bastante *flexível* para incluir também casos de uso de análise de dados offline e machine learning em tempo real.
- O stack também precisa ser *acessível* a partir de qualquer lugar. Ele precisa oferecer suporte a diferentes plataformas, linguagens de programação, ferramentas e padrões abertos.



Google Cloud

Capítulo 1

Por que você ganha vantagem competitiva quando aproveita ao máximo os dados

Todos reconhecem que os dados são importantes, mas poucas empresas conseguem extrair deles insights inovadores sobre os negócios e os clientes. O que significa aproveitar os dados ao máximo? Por que isso é um desafio?

Quando você aproveita os dados ao máximo, eles podem ser usados para a tomada de decisões sobre produtos e operações. Então, faça algumas perguntas a si mesmo. Você sabe se as expectativas dos seus clientes mudaram? Como você usa os dados para melhorar a experiência dos clientes? Quanto ao desafio, como os engenheiros e cientistas de dados estão empregando seu tempo atualmente?

Os dados são cruciais para promover a experiência do usuário e a criação de produtos inovadores, bem como decisões mais abrangentes da estratégia de lançamento no mercado. O bom uso dos dados cria uma grande vantagem competitiva. É por isso que a maioria das empresas e startups de tecnologia sofre uma pressão enorme para ir cada vez mais além: modernizar e operar em escalas cada vez maiores, justificar os custos de dados atuais e futuros e melhorar a maturidade e a tomada de decisões da organização.

No entanto, as questões de acesso, armazenamento, ferramentas inconsistentes, compliance e segurança são desafios que dificultam o processo de análise e descoberta do valor real dos dados.

Talvez você esteja tentando conciliar sistemas legados e novos. Será que os seus dados devem ficar em uma só nuvem? Ou é melhor que sejam distribuídos em várias nuvens? Como modernizar os stacks de análise (antes integrados verticalmente) para trabalhar com plataformas capazes de escalar horizontalmente?

Ou talvez você esteja processando os dados em lotes ou microlotes, e não em tempo real. O sistema de orquestração e a programação resultantes tornam sua arquitetura mais complexa e exigem manutenção em torno da contenção e resiliência. A sobrecarga operacional resultante do gerenciamento e da manutenção de uma arquitetura em lotes é onerosa e compromete a latência dos dados.

Sem ter acesso fácil a todos os dados e sem poder fazer o processamento e a análise deles no momento da entrada, você fica em desvantagem. O conjunto de tecnologias modernas precisa acompanhar a escala dos dados, usar os mais recentes e incorporar e compreender os que não estão estruturados. Além disso, as equipes de análise de dados mais avançadas mudaram o foco da operação para a ação, usando AI/ML para realizar experimentos e operacionalizar processos.

Google Cloud

Capítulo 2

Como fazer com que os dados trabalhem a seu favor para que você possa se concentrar na inovação

O que significa fazer os dados trabalharem a seu favor? Melhorar a experiência do cliente, alcançar clientes novos e aumentar sua receita. Na essência, é uma questão de inovação. Recomendamos dois princípios na escolha de uma plataforma de dados que ajude você a alcançar esses resultados.

Princípio 1: simplicidade e escalabilidade

É provável que você tenha muitos dados à sua disposição no momento. Talvez eles estejam crescendo exponencialmente e você queira manter ou aumentar seu ROI enquanto acompanha o volume. Talvez você esteja prevendo que terá muitos dados no futuro, por exemplo, um terabyte, e projetando seus sistemas para processar esse valor sabendo que, se o crescimento superar essa expectativa, será necessária a migração de todo o sistema. Ou talvez você tenha escolhido um data warehouse capaz de acompanhar a escala do crescimento esperado, mas tem sido difícil gerenciar o aumento das necessidades de processamento.

Sistemas menores costumam ser mais simples. No entanto, você não precisa mais escolher entre um sistema fácil de usar e um altamente escalável. Usar uma arquitetura sem servidor elimina a necessidade de gerenciar clusters e possibilita a utilização de enormes escalas de computação e armazenamento, para que você nunca mais precise se preocupar se o volume dos dados excede sua capacidade técnica.

Para fins de simplicidade e escalabilidade, recomendamos uma plataforma de dados sem servidor. Sugerimos que você desconsidere qualquer opção que exija a instalação de software, o gerenciamento de clusters ou o ajuste de consultas.

Princípio 2: agilidade e custos baixos

Qualquer sistema de gerenciamento de dados que combine computação e armazenamento fará com que você tenha que escalar a computação para lidar com o aumento do volume de dados, mesmo que não precise. Isso pode ser caro e você pode acabar tendo que fazer certas concessões, como armazenar apenas os dados dos últimos 12 meses no warehouse de análise. Você também pode decidir não incluir dados por não ter uma utilidade imediata para eles e descobrir mais tarde que não é possível testar uma hipótese porque eles não estão lá e seria necessário criar um novo pipeline.

Outros sistemas resolvem parte desse problema, permitindo escalar e pagar separadamente pela computação e pelo armazenamento. Ainda assim, você precisa configurar, escalar e otimizar os clusters de forma manual. Para reduzir ao máximo o gerenciamento de infraestrutura, você pode usar um data warehouse sem servidor em várias nuvens com maior confiabilidade, desempenho e proteção de dados integrada, como o BigQuery.

Além do custo e do gerenciamento, você também precisa considerar a agilidade. Quando seus dados mudam, quanto tempo você leva para perceber e agir? Quando uma nova versão de um software ou de uma ferramenta que você usa é lançada, quanto tempo você leva para adotar os novos recursos? O caminho para maior agilidade é escolher ferramentas flexíveis que não exijam monitoramento constante e sejam aplicáveis a uma grande variedade de cargas de trabalho.

As consultas em sistemas como o Redshift precisam ser otimizadas para que sejam eficientes. Isso limita a quantidade de experimentos que podem ser realizados e talvez você acabe extraindo os dados só quando achar que há um problema. As medidas tomadas para lidar com a falta de separação entre computação e armazenamento, bem como a necessidade de otimizar seu data warehouse, acabam limitando você.

Com o BigQuery, você não precisa planejar as consultas com antecedência nem indexar seus conjuntos de dados. A separação do armazenamento e da computação permite que você inclua dados sem se preocupar com o aumento dos custos de consulta, e seus cientistas de dados podem fazer experimentos sem se preocupar com os clusters ou o dimensionamento dos data warehouses para testar novas ideias usando consultas exclusivamente para isso.

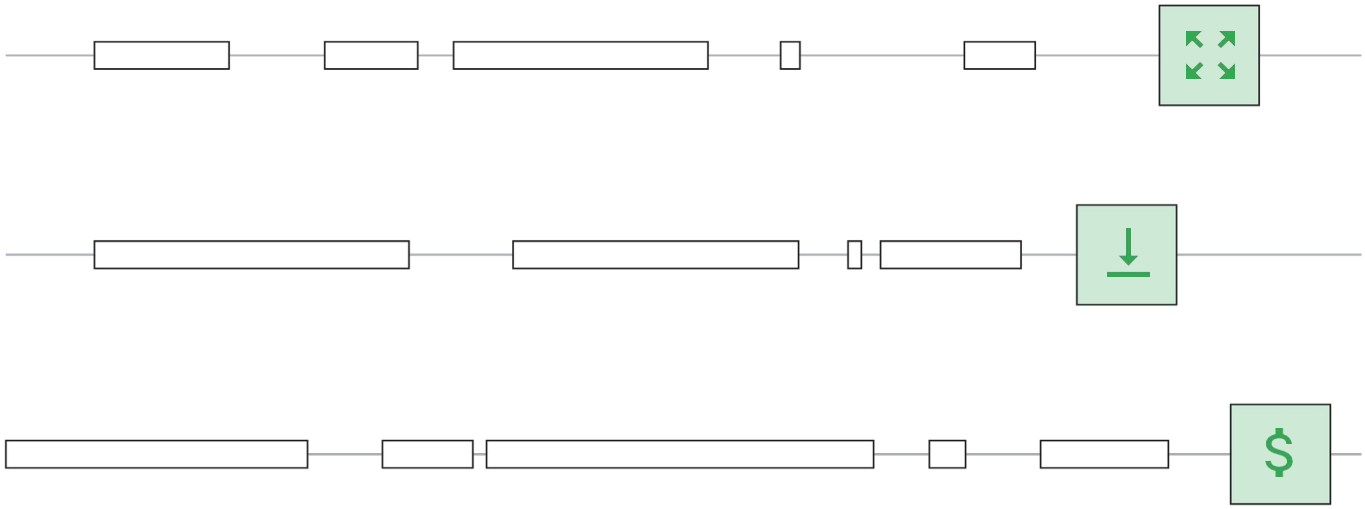
Mostramos como uma plataforma simples, escalável, flexível e econômica coloca você em uma posição que favorece a inovação. Agora vamos ver como seus dados podem ajudar nisso.

Google Cloud

Tome decisões fundamentadas por dados em tempo real

O ritmo de trabalho das empresas não para de acelerar. As expectativas dos clientes também mudaram. Os casos em que seria possível reconciliar uma transação ou aprovar uma devolução em três dias agora precisam de uma resposta imediata. A tomada de decisão mais rápida e pontual gerou uma necessidade maior de usar streaming.

Você quer capturar dados em tempo real e os disponibilizar às suas equipes de negócios para consultas de baixa latência. Você também quer ter certeza de que os pipelines de streaming são escaláveis, resilientes e têm poucas despesas de gerenciamento. Essa é a única maneira de sua equipe reagir em tempo real acompanhando a velocidade dos negócios. Não é de surpreender que o BigQuery tenha suporte nativo para ingestão de dados por streaming e os disponibilize imediatamente para análise usando SQL. Além da API Streaming do BigQuery, que é fácil de usar, o Dataflow possibilita o gerenciamento de cargas de trabalho altas ou sazonais sem gastar demais.



Elimine os silos de dados

Muitas organizações acabam criando silos porque armazenam separadamente os dados de diferentes departamentos e unidades de negócios que cada equipe gera. Isso significa que, sempre que você quiser fazer uma análise que envolva vários departamentos, terá que descobrir como eliminar esses silos, provavelmente executando pipelines de extração (ETL) para obter os dados e colocá-los no data warehouse. No entanto, os departamentos que geraram os dados geralmente têm pouco incentivo para manter os pipelines. Com o tempo, eles ficam desatualizados, e os dados coletados ficam mais obsoletos e menos úteis.

Além dos silos organizacionais, várias empresas têm adotado uma estratégia multinuvem baseada nas preferências de cada departamento, no alinhamento de capacidades e na pressão regulatória. Essas empresas também costumam lidar com a realidade de data lakes legados e investimentos em data warehouse on-premises. O ambiente multinuvem e híbrido de hoje exige um nível maior de sofisticação para gerenciar e acessar dados em silos.

Migrar para um warehouse distribuído com um painel de controle comum, também chamado de malha de dados ou data mesh, aumenta a capacidade de acessar dados de alta qualidade em departamentos, nuvens e sistemas on-premises. Isso pode solucionar problemas comerciais, como desempenho do produto ou comportamento do cliente, e possibilita a consulta dos dados em tempo real.

O BigQuery oferece a base tecnológica para essa malha de dados. Os usuários da organização podem gerenciar, proteger, acessar e compartilhar insights e recursos de dados, independentemente do proprietário. Por exemplo, é possível levar todos os seus dados para o BigQuery e proporcionar funções reutilizáveis, visualizações materializadas e até mesmo a capacidade de treinar modelos de ML sem a necessidade de movê-los. Isso significa que até mesmo especialistas da área sem conhecimento técnico, assim como parceiros e fornecedores autorizados, podem acessar e usar o SQL facilmente para consultar os dados com ferramentas conhecidas, como planilhas e painéis.

A ilustração do modelo "hub-and-spoke" é bem adequada neste caso. O BigQuery é o centro (hub) onde estão os dados. Os spokes são ferramentas de relatórios, painéis, modelos de ML, aplicativos da Web, sistemas de recomendação e outros. Todos leem dados do BigQuery sem criar cópias. O Looker, por exemplo, ajuda a visualizar os dados e fazer a integração deles ao fluxo de trabalho diário dos usuários. Essa abordagem permite aprimorar a usabilidade, segurança e qualidade dos dados, tudo de uma vez.

Simplifique o acesso a todos os dados

Antes, a melhor forma de processar dados semiestruturados ou não estruturados era usando data lakes. Já os data warehouses eram melhores para dados estruturados. Essa separação criou silos tecnológicos que dificultaram o cruzamento dos formatos. Os dados eram todos armazenados em data lakes, que eram mais baratos e fáceis de gerenciar. Depois, eles eram movidos para um warehouse onde ferramentas de análise eram usadas na extração de insights.

O modelo cada vez mais conhecido de "lake house" combina esses dois mundos em um ambiente unificado para todos os tipos de dados. O BigQuery pode ser usado tanto como um data warehouse quanto um data lake. A API Storage do BigQuery permite acessar direto o armazenamento para processar cargas de trabalho geralmente associadas a data lakes. Como os dados podem ser armazenados em uma única fonte da verdade no BigQuery, não é necessário criar nem manter muitas cópias.
Em vez disso, o processamento downstream pode ser realizado por transformações SQL que são armazenadas em visualizações lógicas sem precisar mover os dados.

A facilidade de uso é importante. Se o resultado da sua consulta aparecer em 30 segundos, e não em 30 minutos ou 3 horas, provavelmente você conseguirá usar melhor os dados na tomada de decisões.

Use AI/ML para realizar experimentos com rapidez e operacionalizar cargas de trabalho

Seus cientistas de dados conseguem realizar experimentos com agilidade?
É provável que tenham que parar o desenvolvimento e operacionalizar modelos para avaliar os experimentos com usuários reais. Os cientistas desenvolvem e fazem iterações de um modelo usando dados históricos antes de entregá-lo ao departamento de engenharia, que o reescreve completamente para incorporar ao sistema de produção e realizar testes A/B. Em seguida, há um período de espera, a iteração com base no modelo e o envio de volta para produção. Esse ciclo envolve muitas interrupções e reescritas no código, e a coordenação exigida entre as equipes geralmente ocasiona erros. Seus cientistas de dados não fazem todos os experimentos possíveis, porque isso levaria muito tempo. Assim fica difícil prever quanto tempo um projeto pode levar e se será bem-sucedido, além do período necessário para ser usado rotineiramente. Para superar esse problema, você precisa fornecer aos seus cientistas de dados ferramentas avançadas, porém familiares. Com o Vertex AI Workbench, eles trabalham com eficiência em Jupyter notebooks, mas o treinamento, a experimentação e a implantação são rápidos.

Se você quiser se destacar com base nos dados, precisa extrair o maior valor possível do que coleta. Para isso, é necessário que suas equipes de cientistas de dados sejam as mais produtivas possíveis e não percam a chance de criar modelos, porque até mesmo as coisas simples podem levar muito tempo ou serem muito difíceis.

A qualidade dos modelos pré-criados e com pouco código é crucial. O AutoML na Vertex AI disponibiliza os melhores modelos de AI em um ambiente sem código, o que agiliza o desenvolvimento de benchmarks e a priorização. Usando modelos pré-criados, como o Entity Extraction ou o Vertex AI Matching Engine, nos seus próprios dados, você acelera significativamente a criação de valor a partir dos dados e não se limita apenas à classificação ou à regressão.

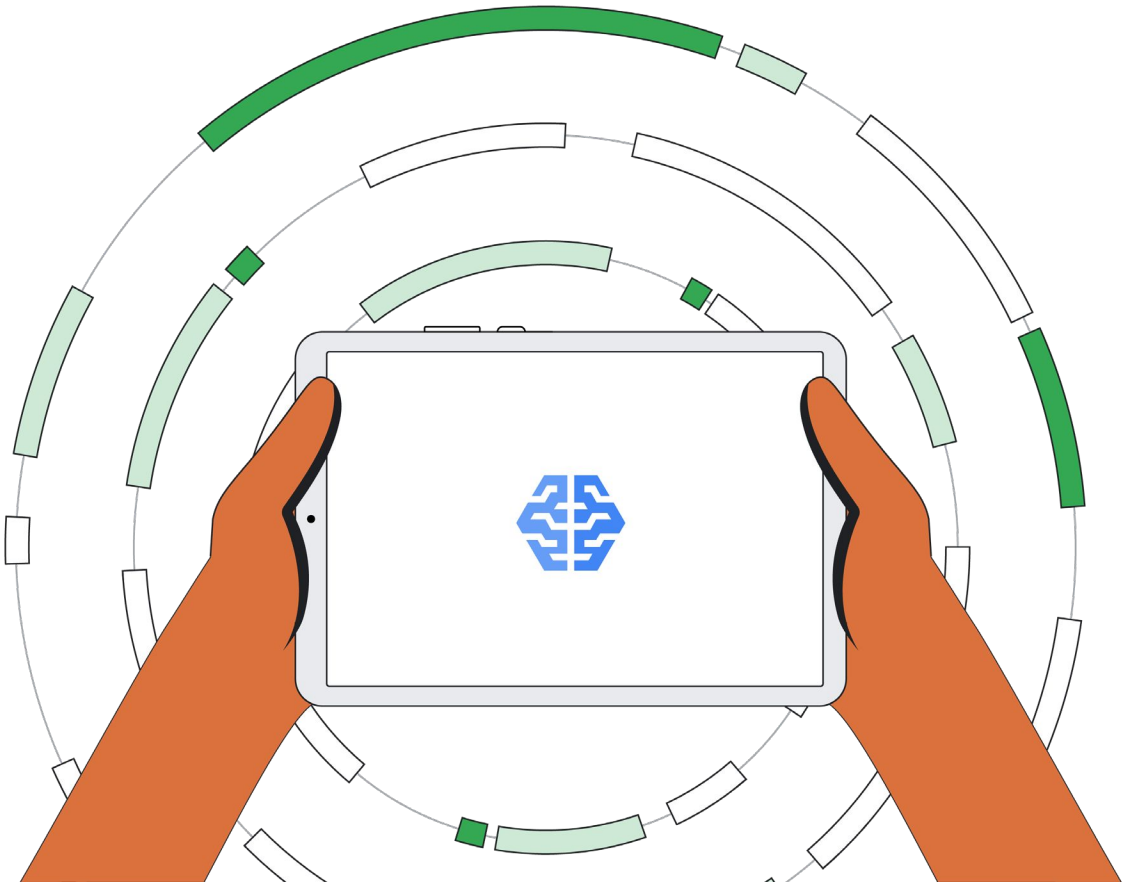
A chave para manter a agilidade dos seus dados é sempre realizar experimentos completos e frequentes. O Vertex AI Pipelines oferece um histórico de experimentos que permite ver o que foi feito e comparar com benchmarks e endpoints, e realizar testes A/B com modelos de sombra. Como o código é conteinerizado, ele pode ser usado em sistemas de desenvolvimento e produção. Os cientistas de dados trabalham em Python, e a equipe de engenharia de produção trabalha com contêineres totalmente encapsulados. Ambas as equipes podem padronizar operacionalizando os modelos com a Vertex AI Prediction, e você pode agir rapidamente.

Os especialistas da área normalmente usam o BigQuery ML para testar a viabilidade de uma ideia treinando modelos personalizados trabalhando com SQL sem precisar ter uma experiência extra com ferramentas tradicionais de ciência de dados. Isso significa que é possível fazer testes em um sistema semelhante ao de produção e realizar estudos de viabilidade em questão de dias, em vez de meses. O modelo do BigQuery ML pode ser implantado na Vertex AI para você aproveitar todas as vantagens abordadas. É possível usar o Looker para criar modelos consistentes com base em todos os seus dados e usar o LookML para consultas, o que significa que todos na organização podem criar relatórios e painéis fáceis de ler para explorar padrões de dados.

Para aumentar o valor real da produção, os sistemas precisam ser capazes de ingerir, processar e exibir dados. Além disso, o machine learning precisa promover serviços personalizados em tempo real com base no contexto do cliente. No entanto, para um aplicativo de produção que é executado de forma contínua, é necessário que os modelos sejam sempre treinados, implantados e verificados para garantir a segurança. Os dados de entrada exigem pré-processamento e validação para garantir que não haja problemas de qualidade e, em seguida, vem a engenharia de atributos e o treinamento de modelos com o ajuste de hiperparâmetros.

A integração da ciência de dados e do machine learning é essencial para orquestrar e gerenciar facilmente esses fluxos de trabalho de ML multifásicos, para que sejam executados de maneira confiável e repetida. Com as ferramentas e os fluxos automatizados do MLOps, é possível ter entregas rápidas e contínuas, além de simplificar o gerenciamento de modelos para a produção. Como só existe um fluxo de trabalho e vocabulário para todos os nossos produtos de AI, independentemente da camada de abstração, é fácil passar de um modelo personalizado para o AutoML, porque eles têm o mesmo formato e base técnica.

Por exemplo, e se você quiser aplicar a detecção de anomalias a fluxos de dados ilimitados e em tempo real para combater fraudes? Com a abordagem certa, poderá gerar um fluxo de dados de amostra para simular o tráfego de rede comum e transferir para o Pub/Sub e, depois, criar e treinar um modelo de detecção de anomalias no BigQuery usando a clusterização k-means do BigQuery ML após mascarar as informações de identificação pessoal (PII) com DLP. Em seguida, você pode aplicar o modelo aos dados ao vivo para detecção em tempo real com o Dataflow, além do Looker, para criar um painel, alertas e ações para tratar dos eventos identificados.



Capítulo 3

Por que é importante escolher uma opção de data warehouse eficiente

Falamos sobre o BigQuery e o Redshift, mas essas não são as únicas opções de data warehouse disponíveis. Há outros produtos de análise de dados (como o Snowflake e o Databricks) que funcionam nas três nuvens principais. Se você escolher o BigQuery, o vínculo com essa nuvem vai ser um problema?

A primeira coisa a ser observada é que, com o BigQuery, você não se limita a analisar apenas os dados armazenados no Google Cloud. Com o BigQuery Omni, é possível consultar seus dados no Amazon S3 e no Armazenamento de Blobs do Azure usando o console do Google Cloud.

No entanto, a realidade é que, se você usar o Snowflake ou o Databricks, os custos de migração do AWS para o Google Cloud ou vice-versa serão mais baixos. E os custos de migração para outro data warehouse? E se você quiser migrar do Snowflake para o BigQuery ou do Databricks para EMR? O custo de migração continua existindo; só muda o cenário.

Como haverá custos de migração em qualquer cenário, você precisa escolher a ferramenta ou a plataforma mais adequada no longo prazo. Você deve escolher uma plataforma com base nos recursos que a diferenciam, no custo atual e na rapidez com que ela vai gerar inovação no futuro. Quando você opta pelo Snowflake, está apostando que uma empresa focada em data warehouse vai oferecer inovações mais ágeis nessa área. Ao escolher o BigQuery, você conta com uma empresa conhecida por inventar muitas tecnologias de dados e AI para continuar inovando na plataforma.

Acreditamos que uma plataforma inovadora e bem integrada potencializa melhor o efeito volante da inovação. Quando uma oferta de serviço gerenciado como o Google Kubernetes Engine (GKE) faz com que as imagens de contêiner sejam carregadas mais rapidamente, o Spark sem servidor funciona melhor e, como é capaz de operar dados no BigQuery, o BigQuery agrega mais valor para você. O volante gira mais rápido quando você aposta em uma plataforma, e não em produtos separados.

Capítulo 4

Como fazer sua jornada de migração de dados com segurança

Quanto tempo dura a migração de dados? Seis meses? Dois anos? Quanto esforço ela exige? Vale a pena?

Se você está migrando de uma nuvem para outra, provavelmente será mais fácil do que migrar do ambiente on-prem para a nuvem devido ao fato de que geralmente a tecnologia local é muito mais complexa. Apesar disso, procure se concentrar na sua meta, que provavelmente deve envolver a questão: "Com que rapidez vou conseguir inovar?".

Pense em todas as coisas que você quer fazer e que não está fazendo hoje. Em seguida, configure novos projetos e transfira os dados necessários para fazer o que quiser. Podemos ajudar você a criar esses novos casos de uso e espelhar as fontes de dados de que precisa. Você vai operar em um ambiente híbrido temporariamente, no qual muitos casos de uso são executados on-premises, porém são baseados em dados espelhados em tempo real ou em lote desse ambiente ou de outro provedor de nuvem.

A segunda consideração que você deve fazer é relacionada ao custo. Vamos tomar como exemplo as instâncias extremamente caras do Teradata que você está executando. Ao migrar para o BigQuery, notamos que os clientes reduziram os custos pela metade, e essas migrações são muito mais fáceis do que antes, graças às ferramentas de avaliação automatizadas e aos transpiladores automatizados de SQL que convertem a grande maioria dos scripts. Temos como virtualizar as coisas para que seus clientes pensem que estão se comunicando com o Teradata, quando na verdade estão falando com o BigQuery. Há muitas maneiras de ajudarmos você a migrar sem precisar desativar todos os recursos. Essas ferramentas de migração permitem que você pare de usar as cargas de trabalho dispendiosas do Teradata e do Hadoop.

A terceira consideração é analisar seus sistemas de ERP, como SAP, Salesforce e Oracle. Se você quer otimizar sua cadeia de suprimentos, pontuar leads ou detectar fraudes, é importante conectar suas cargas de trabalho de análise aos sistemas de ERP. Há conectores externos que podemos usar para coletar dados desses sistemas que, por sua vez, podem ser empregados para criar casos de uso modernos baseados em AI com esses dados da nuvem.

A ordem para fazer essas coisas depende da situação. Se sua empresa for uma startup, comece com a inovação, a otimização de custos e, por fim, aproveite os pipelines e conectores. Se ela depender muito da cadeia de suprimentos, comece com os conectores de ERP. Independentemente da ordem, você vai ver que consegue migrar uma quantidade considerável dos seus valiosos dados para a nuvem. Depois, veja o que restou e decida se vale a pena migrar tudo. Muitas vezes, a resposta é não. Depois de migrar 70% a 80% das cargas de trabalho realmente necessárias, será preciso tomar decisões difíceis. Vale a pena migrar os outros 20% a 30%? Não seria melhor reescrever o código ou fazer a tarefa de outra forma? Se você começar a migrar tudo para a nuvem do jeito que está, vai acabar repetindo no ambiente de nuvem as despesas com tecnologia que tinha on-premises, em vez de manter o foco no valor dos dados.