Google BigQuery e Tableau: práticas recomendadas

Introdução

O Tableau e o Google BigQuery permitem que as pessoas analisem grandes volumes de dados e obtenham respostas com rapidez por meio de uma interface visual fácil de usar. Usando as duas ferramentas juntas, você pode:

- Colocar toda a eficiência do Google BigQuery nas mãos de usuários comuns para análises rápidas e interativas.
- Analisar bilhões de linhas em questão de segundos usando ferramentas de análise visual sem escrever uma só linha de código e sem a necessidade de gerenciamento no servidor.
- Criar em questão de minutos painéis incríveis que se conectam aos dados do Google BigQuery e mantêm sua organização atualizada.
- Compartilhar relatórios e informações na Web usando o Tableau Server e o Tableau Online para permitir que qualquer pessoa se conecte a partir de qualquer dispositivo.
- Combinar a agilidade da nuvem do Google BigQuery com a altíssima velocidade do Tableau para identificar o valor de projetos com mais rapidez.

A otimização das duas tecnologias juntas aumentará consideravelmente o desempenho, reduzirá os ciclos de desenvolvimento e ajudará os usuários e as organizações a terem mais sucesso. Neste documento, abordaremos técnicas para otimizar a modelagem de dados e a formação de consultas para maximizar a capacidade de resposta das visualizações. Também abordaremos técnicas para obter o melhor custo/benefício ao usar o Tableau e o BigQuery juntos.

Autores

Pierce Young, gerente de produtos, Tableau

Vaidy Krishnan, gerente sênior de produtos, Tableau

Riley Maris, especialista sênior em marketing de produtos, Tableau

Babu Prasad Elumala, engenheiro de soluções, Google

Seth Hollyman, gerente de programas técnicos, Google

Tino Tereshko, engenheiro de soluções empresariais, Google

Mike Graboski, engenheiro de soluções, Google

Sumário

Visão geral da tecnologia

Google BigQuery

O BigQuery pode processar petabytes de dados em questão de segundos em SQL simples sem necessidade de ajustes nem de habilidades especiais. Integrado ao sistema Dremel, a tecnologia revolucionária do Google para analisar conjuntos de dados de grande volume, o BigQuery oferece um nível de desempenho que as grandes empresas antes precisavam pagar milhões para obter: tudo isso a um preço de centavos por gigabyte.

O BigQuery é um data warehouse mais adequado para executar consultas SQL em conjuntos de dados de grande volume, estruturados e semiestruturados. Exemplos de casos de uso e conjuntos de dados incluem:

- Análises ad hoc
- Logs da Web
- Logs de máquinas/servidores
- Conjuntos de dados da Internet das Coisas
- Comportamento de clientes de comércio eletrônico
- Dados de aplicativos móveis
- Análises no setor varejista
- Telemetria em jogos
- Dados do Google Analytics Premium
- Qualquer conjunto de dados no qual um RDBMS tradicional leva minutos (ou horas) para executar uma consulta em lote

O BigQuery dispensa totalmente a intervenção da equipe de operações e é integrado ao Google Cloud Platform. Diferentemente de outras soluções de análise baseadas na nuvem, o BigQuery não exige que você provisione um cluster de servidores antes. Os clusters de processamento são dimensionados e provisionados pelo BigQuery no momento da execução.

O BigQuery adiciona capacidade de processamento automaticamente à medida que o seu volume de dados aumenta. No entanto, você paga o mesmo preço por gigabyte.

SQL herdado x SQL padrão

O Google BigQuery atualizou suas APIs para utilizar o SQL padrão além do BigQuery SQL (agora chamado de SQL herdado), e o Tableau atualizou seu conector do Google BigQuery para oferecer suporte a essa mudança para o SQL padrão. O SQL padrão oferece benefícios aos usuários do BigQuery, incluindo Expressões de nível de detalhe, validação de metadados mais rápida e a opção de selecionar um projeto de faturamento com a sua conexão. Este guia foi elaborado com o SQL padrão em mente.

Para obter mais informações sobre como migrar do SQL herdado para o SQL padrão, consulte nosso guia da Ajuda on-line que explica como migrar do SQL herdado (em inglês).

Tableau

O Tableau ajuda as pessoas a ver e a entender os dados. Nossa plataforma de análise moderna, que tem como base uma tecnologia desenvolvida na Universidade Stanford, coloca o poder dos dados ao alcance de todos. Isso permite que um amplo grupo de usuários interaja com seus dados, faça perguntas, resolva problemas, compartilhe informações e agregue um valor transformador. Independentemente de estarem ou não familiarizadas com ferramentas de BI, as pessoas aprendem rapidamente a usar o Tableau para criar e explorar visualizações interativas sofisticadas e painéis avançados em uma interface intuitiva de arrastar e soltar.

Mais recentemente, ampliamos a funcionalidade da plataforma do Tableau para incluir recursos de preparação de dados visuais, diretos e inteligentes, bem como a possibilidade de consultar fontes de dados publicadas usando a linguagem natural.

Otimizações nativas do Tableau

Conector de fonte de dados: o Tableau possui um conector nativo otimizado para o Google BigQuery que oferece conectividade aos dados em tempo real e extrações na memória. A combinação de dados do Tableau permite que os usuários combinem dados do BigQuery com dados de qualquer uma de nossas mais de 67 fontes de dados compatíveis. Para visualizações publicadas na nuvem com o Tableau Server ou o Tableau Online, é possível manter a conectividade direta com o Google BigQuery.

Consultas paralelas: o Tableau aproveita os recursos do Google BigQuery e de outras fontes de dados para executar várias consultas ao mesmo tempo, totalizando até 16 consultas simultâneas. Lotes de consultas independentes e consolidadas serão agrupados e enviados ao BigQuery se o resultado ainda não estiver armazenado em cache. Os usuários observarão um aumento considerável no desempenho devido às consultas paralelas, graças à arquitetura de escalabilidade horizontal do BigQuery.

Fusão de consultas: o Tableau recebe e, quando possível, funde várias consultas de pastas de trabalho e painéis, reduzindo o número de consultas enviadas ao BigQuery. Primeiro, o Tableau identifica consultas semelhantes, excluindo as diferenças nas colunas retornadas. Em seguida, ele combina consultas cujas diferenças sejam somente o nível de agregação ou um cálculo do usuário.

Cache de consultas externo: se a fonte de dados subjacente não tiver mudado desde a última vez que você executou determinada consulta, o Tableau automaticamente lerá o cache de consultas salvo anteriormente, oferecendo tempos de carregamento quase instantâneos.

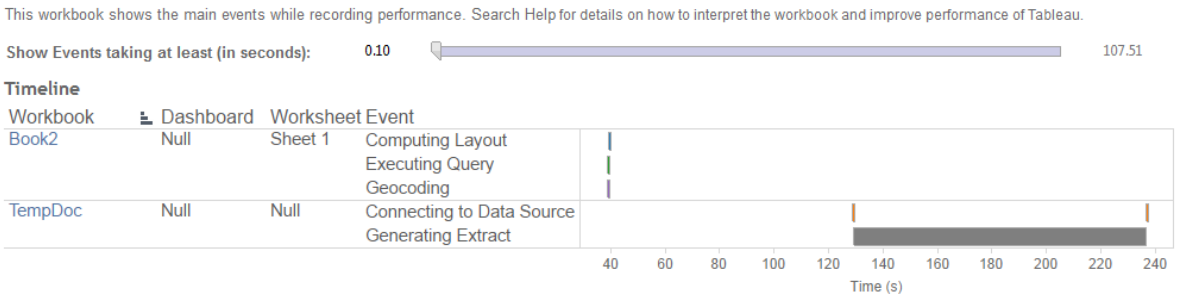
Conexões sob demanda no Tableau Desktop: quando você abre uma pasta de trabalho publicada, o Tableau Desktop apenas se conecta às fontes de dados necessárias para exibir os dados da planilha atual. Em outras palavras, seus dados serão exibidos com mais rapidez.

Práticas recomendadas de desempenho: Tableau

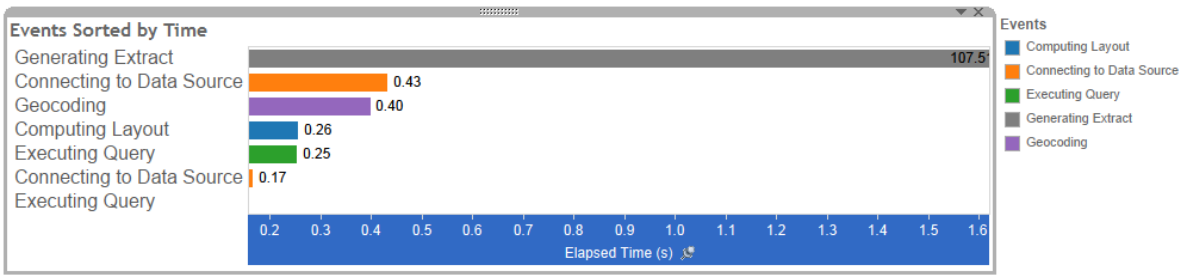
Antes de entrarmos no tópico de ferramentas adicionais e configurações personalizadas, nossa primeira recomendação é que você continue atualizando sua implantação do Tableau quando puder. Assim, você pode aproveitar as melhorias de desempenho que incorporamos continuamente nas versões mais recentes do produto.

Registrador de desempenho

O Registrador de desempenho é uma ferramenta integrada eficiente que permite identificar consultas lentas e otimizar as pastas de trabalho para obter máximo desempenho. Ele faz isso monitorando o tempo que uma pasta de trabalho individual leva para executar uma consulta e processar o layout. Quando o usuário passa o mouse sobre uma das barras verdes abaixo, a consulta que está sendo gerada no BigQuery é exibida. Após identificar uma consulta lenta, você geralmente pode resolver o problema de desempenho revisando seu modelo de dados.



Na exibição Linha do tempo, as colunas Pasta de trabalho, Painel e Planilha identificam o contexto de eventos.



Eventos com durações mais longas podem ajudar a identificar onde você deve investigar primeiro caso queira acelerar sua pasta de trabalho.

Para obter mais informações sobre como criar ou interpretar um registro de desempenho, consulte os seguintes artigos:

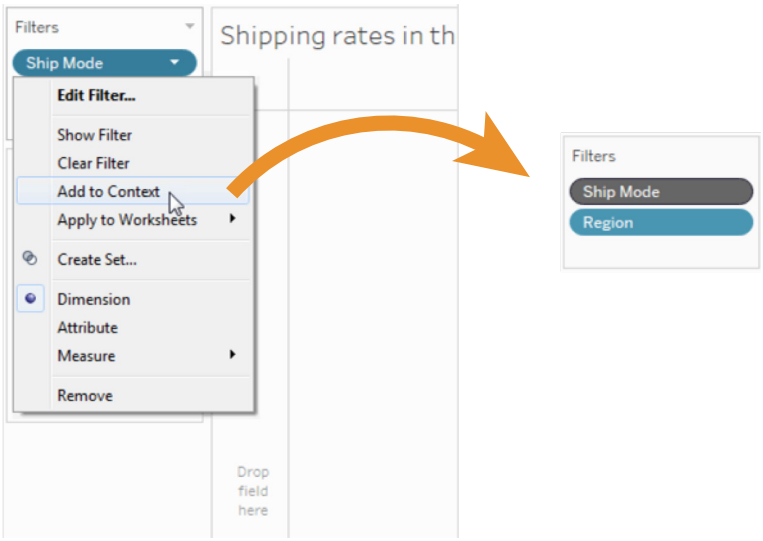
Registrador de desempenho no Tableau Desktop (criar)
Registrador de desempenho no Tableau Server (interpretar)

Filtros de contexto

Se você for aplicar filtros a uma fonte de dados grande, poderá melhorar o desempenho configurando filtros de contexto. Um filtro de contexto é aplicado à fonte de dados primeiro, para que filtros adicionais sejam aplicados somente aos registros resultantes. Essa sequência evita que cada filtro seja aplicado a cada registro na fonte de dados.

Se você for aplicar filtros que reduzem consideravelmente o tamanho do conjunto de dados e for usar esses filtros para muitas exibições de dados, defina esses filtros como filtros de contexto.

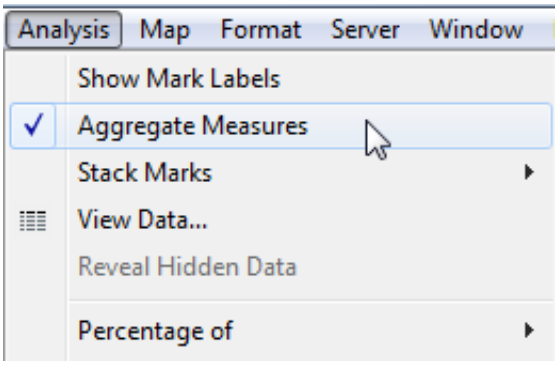
Para obter mais informações, consulte nosso guia da Ajuda on-line que explica como melhorar o desempenho de exibições com filtros de contexto.



Você pode definir um ou mais filtros de contexto para melhorar o desempenho.

Agregar medidas

Se as exibições que você cria são lentas, verifique se está trabalhando com medidas agregadas em vez de medidas desagregadas. Quando as exibições são lentas, geralmente significa que você está tentando visualizar muitas linhas de dados de uma só vez. Você pode reduzir o número de linhas agregando os dados.



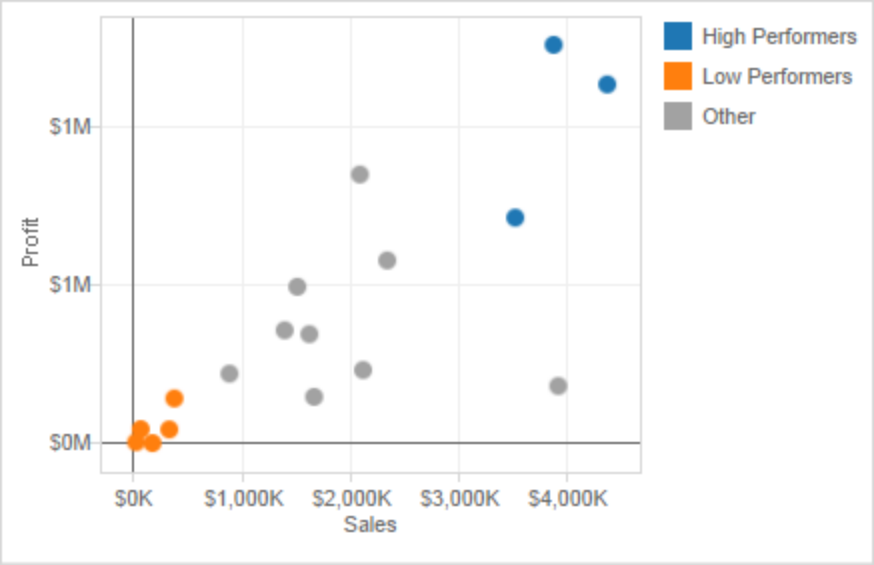
Veja se as medidas são agregadas no menu Análise. Você também pode definir agregações padrão para medidas.

Para obter mais informações, consulte nosso guia da Ajuda on-line sobre agregação de dados.

Conjuntos e grupos

Se você deseja filtrar uma dimensão para remover membros com base em um intervalo de valores de medida, crie um conjunto em vez de usar um filtro quantitativo. Por exemplo, você pode criar um conjunto que retorna apenas os 50 primeiros itens em uma dimensão, em vez de todos os itens.

Ao criar um grupo a partir de uma seleção, inclua somente as colunas relevantes. Cada coluna adicional incluída no conjunto resultará em queda do desempenho.



Ao criar grupos no Tableau, você tem a opção de agrupar todos os membros restantes em um grupo chamado Outros.

Para obter mais informações, consulte nosso guia da Ajuda on-line sobre como criar conjuntos *e nosso* guia da Ajuda on-line sobre como criar grupos.

Adicionar filtros primeiro

Se você estiver trabalhando com uma fonte de dados grande e tiver as atualizações automáticas desativadas, poderá criar uma consulta muito lenta ao adicionar filtros à exibição. Em vez de criar a exibição e depois especificar filtros, você deve primeiro definir os filtros e depois arrastar campos para a exibição. Assim, quando você executar a atualização ou ativar as atualizações automáticas, os filtros serão avaliados primeiro.

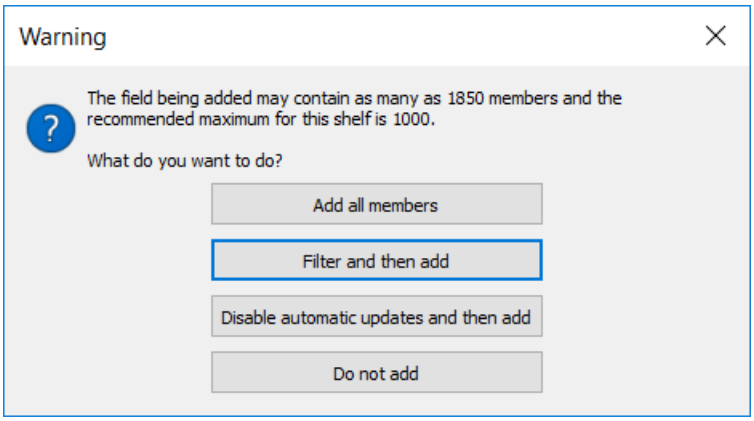
Desativar atualizações automáticas

Quando você coloca um campo em uma divisória, o Tableau gera a exibição consultando automaticamente a fonte de dados. Se você estiver criando uma exibição de dados densa, as consultas poderão ser demoradas e poderão degradar consideravelmente o desempenho do sistema. Nesse caso, você pode instruir o Tableau a desativar consultas enquanto cria a exibição. Em seguida, você pode ativar novamente as consultas quando estiver pronto para ver o resultado.

Para obter mais informações, consulte nosso guia da Ajuda on-line sobre atualizações automáticas e desempenho.

Procurar avisos

O Tableau exibe uma caixa de diálogo de aviso de desempenho quando você tenta colocar uma dimensão grande (com muitos membros) em qualquer divisória. A caixa de diálogo oferece quatro opções, conforme mostrado na figura abaixo. Se você optar por adicionar todos os membros, poderá haver uma queda significativa no desempenho.



Quando você colocar uma dimensão grande em uma divisória, o Tableau avisará sobre o risco de queda de desempenho.

Otimizar consultas paralelas

Você pode usar atributos de personalização para melhorar o desempenho de conjuntos de resultados grandes retornados pelo BigQuery para o Tableau Online ou o Tableau Server, assim como no Tableau Desktop, configurando consultas paralelas. Esses atributos de personalização podem ser incluídos em sua pasta de trabalho ou fonte de dados publicada, desde que você especifique os atributos antes de publicar a pasta de trabalho ou a fonte de dados no Tableau Online ou no Tableau Server.

Para obter mais informações, consulte "Usar atributos de personalização para melhorar o desempenho da consulta" em nosso guia da Ajuda on-line sobre o Google BigQuery.

Práticas recomendadas de custo e desempenho: Google BigQuery

Para garantir um desempenho alto nas consultas e reduzir os custos, evite usar tabelas federadas com dados armazenados em uma fonte de dados externa, como o Google Cloud Storage. Em situações como essa, se você deseja realizar consultas iterativas no conjunto de dados, use a API de consulta para materializar os dados no BigQuery (independentemente do Tableau) e permitir um desempenho de consulta alto no conjunto de dados com o Tableau.

Desnormalizar e pré-unir

O BigQuery é compatível com uniões extremamente grandes, e o desempenho de união é excelente. No entanto, o BigQuery é um datastore colunar e o desempenho máximo é atingido em conjuntos de dados desnormalizados.

Um dos benefícios da nuvem é a possibilidade de separar os recursos de armazenamento e processamento, permitindo que os usuários escalonem e paguem por cada recurso individualmente. Como o armazenamento do BigQuery tem um custo muito baixo e é altamente escalonável, geralmente é recomendável desnormalizar e pré-unir os conjuntos de dados em tabelas homogêneas. Basicamente, isso significa que você usará menos recursos de processamento e mais recursos de armazenamento (estes últimos oferecem maior desempenho e são mais econômicos). Como o BigQuery é um repositório colunar e pode compactar os dados de forma eficaz, reduzir o consumo de processamento em troca de um aumento no consumo de armazenamento é uma boa opção e provavelmente custará menos.

O BigQuery é uma excelente ferramenta de ETL, permitindo que você execute funis e transformações enormes com rapidez e eficiência. Ative a opção "Permitir resultados grandes" ao materializar conjuntos de dados com mais de 128 MB.

Para obter mais informações sobre como preparar dados para carregamento e como consultar dados usando a linguagem SQL do BigQuery, consulte os seguintes artigos:

Como fazer o carregamento de dados desnormalizados, aninhados e repetidos (em inglês)
Como gravar resultados de consulta extensos (em inglês)

Segmentar tabelas por data

Dividir uma tabela em partições menores ("segmentos") pode ajudar a simplificar o gerenciamento de dados, a melhorar o desempenho das consultas e a reduzir os custos. Além disso, o BigQuery permite o clustering em uma tabela particionada, o que é útil quando os dados já estão particionados em uma coluna de data ou de carimbo de data e hora ou quando você tem filtros ou agregações de colunas específicas em suas consultas.

Alguns dados são naturalmente adequados para serem particionados por data: por exemplo, dados de logs ou qualquer dado cujos registros incluam um carimbo de data e hora que aumenta gradualmente. Nesse caso, segmente suas tabelas do BigQuery por data e inclua a data no nome da tabela. Para aproveitar essa possibilidade, você precisaria usar SQL personalizado no Tableau.

Para obter mais informações, consulte nosso guia da Ajuda on-line sobre como se conectar a uma consulta SQL personalizada.

Por exemplo, atribua às suas tabelas um nome como: mytable_20170501, mytable_20170502, etc.

Quando você quiser executar uma consulta que filtre por data, use a função de tabela curinga do BigQuery:

```
SELECT
 name

FROM
 `myProject.myDataSet.mytable_*'

WHERE
 age >= 35
```

O exemplo acima automaticamente incluirá todas as tabelas com o prefixo mytable_.

Para usar um curinga, suas tabelas devem ser nomeadas de acordo com este padrão: [qualquer prefixo] AAAAMMDD.

Outros sistemas de banco de dados usam a segmentação para melhorar o desempenho. Na realidade, a segmentação por data tem uma diferença de desempenho mínima no BigQuery, mas o principal fator aqui é o custo. Como menos dados são processados, você paga menos por consulta.

Importante: se você decidir segmentar por minuto, muitas partições serão criadas e isso afetará diretamente o desempenho. Tenha cuidado para não segmentar demais de uma só vez. Qualquer segmentação mais abrangente do que por dia é aceitável.

Para obter mais informações sobre a segmentação, consulte os seguintes artigos:

Introdução a tabelas particionadas (em inglês)
Introdução às tabelas em cluster (em inglês)
Como consultar várias tabelas usando uma tabela curinga (em inglês)

Especificar uma tabela de destino ao executar várias consultas semelhantes

Embora o cache de consultas seja útil ao executar muitas consultas idênticas, ele não ajudará se você estiver executando consultas semelhantes, mas ligeiramente diferentes (por exemplo, só mudam os valores em uma cláusula WHERE entre execuções de consulta). Nesse caso, execute uma consulta na tabela de origem e grave os registros que você consultará repetidamente em uma nova tabela de destino. Em seguida, execute consultas na nova tabela de destino que criou.

Por exemplo, digamos que você pretenda executar três consultas com três cláusulas WHERE diferentes:

```
WHERE col1 = “a”
WHERE col1 = “b”
WHERE col1 = “c”
```

Execute uma consulta na sua tabela de origem e grave os registros resultantes em uma tabela de destino:

```
SELECT col1
FROM source
WHERE col1 = “a” OR col1 = “b” OR col1 = “c”
```

Ao usar “OR” para vincular as cláusulas WHERE, capturamos todos os registros relevantes. Nossa tabela de destino possivelmente será muito menor do que a tabela de origem inicial. Como o BigQuery cobra pela quantidade de dados processados em uma consulta, você economizará dinheiro ao executar as consultas subsequentes na nova tabela de destino em vez de executá-las diretamente na tabela de origem. Lembre-se de apagar essas tabelas no futuro para não acumular custos com seu armazenamento.

Usando o Tableau para visualizar resultados de modelos do Google BigQuery ML

O BigQuery ML permite que os usuários utilizem a tecnologia de aprendizado de máquina incorporada para treinar modelos com base nos dados armazenados no BigQuery. No entanto, como acontece quando se trabalha com qualquer outro tipo de dados, consultar diretamente um banco de dados nem sempre é o método ideal para explorar os resultados do seu modelo.

Com seu conector do BigQuery, o Tableau permite que você manipule os resultados de seus modelos preditivos com facilidade, de uma maneira que possibilita uma compreensão intuitiva dos dados. Além disso, o Tableau oferece aos usuários uma forma fácil de compartilhar seu modelo e seus resultados com outras pessoas para que elas possam aproveitar todo o seu trabalho.

Para ver um exemplo, leia sobre como usar o BigQuery ML com o Tableau para prever os preços de moradia.

Embora o SQL personalizado seja necessário neste exercício para invocar um algoritmo de aprendizado de máquina no Google BigQuery ML, há considerações de desempenho às quais você deve atentar ao usar o SQL personalizado em vez das conexões nativas do Tableau em outras situações. Quando possível, aproveite as conexões de fontes de dados nativas do Tableau para garantir um desempenho ideal.

Para obter mais informações sobre como criar e executar modelos de aprendizado de máquina usando consultas SQL padrão, consulte o guia da Ajuda on-line sobre o Google Big Query ML (em inglês).

Estudo de caso: Principais dicas da zulily para a análise de autoatendimento com o Tableau e o BigQuery

A zulily é uma empresa de comércio eletrônico em rápido crescimento que criou uma plataforma de Big Data usando o Google BigQuery como data warehouse e o Tableau para acesso aos dados e análise visual. A integração do BigQuery e do Tableau permite que a análise acelere os processos de adquirir, processar e usar dados para criar relatórios e modelos sem a intervenção da TI nas atividades diárias. Além disso, os usuários corporativos têm acesso em tempo real aos dados importantes necessários para tomar decisões com rapidez, evitando que os analistas precisem gerar informações básicas.

Estas são algumas das práticas recomendadas da zulily:

Reduza a latência usando o Tableau Server no Google Compute Engine: em vez de um modelo tradicional em que as regiões são VPCs separadas, você pode usar a estrutura privada do Google sem acessar a Internet e sem configuração adicional. Isso também permite que você dimensione suas implantações com precisão, sem a necessidade de superprovisionamento.

Use fontes federadas e aponte o Tableau para o BigQuery: para dados no Google Cloud, você deve aproveitar a capacidade do BigQuery de consultar fontes de dados externas e tratar o BigQuery como seu lago de dados. Em alguns casos, é possível reduzir a quantidade de dados que precisam ser transmitidos pela rede e exibidos no Tableau para análise.

Processe conjuntos de dados grandes com uma conexão em tempo real no BigQuery: aproveite a capacidade do BigQuery de processar conjuntos de dados grandes e envie apenas os resultados pela rede. Defina sua conexão padrão do Tableau com o BigQuery como "em tempo real", a menos que tenha um motivo específico para extrair os dados.

Para obter mais informações e conferir a lista completa com todas as dez dicas, leia a série de duas partes do blog:

Parte 1: Por que a zulily criou uma plataforma de análises de marketing de autoatendimento com o Tableau e o Google BigQuery

Parte 2: 10 principais dicas da zulily para a análise de autoatendimento com o Google BigQuery e o Tableau Desktop (em inglês)

Conclusão

Ao aplicar práticas recomendadas, os usuários corporativos e os analistas de dados poderão maximizar o desempenho e a capacidade de resposta de visualizações do Tableau integradas ao Google BigQuery. Quando essas tecnologias são combinadas, os usuários realmente podem visualizar bilhões de colunas de dados na velocidade do pensamento.

Sobre a Tableau

O Tableau é uma plataforma de business intelligence visual completa, fácil de usar e voltada para empresas. Ele ajuda as pessoas a ver e a entender os dados com análises de autoatendimento ágeis em qualquer escala. Seja na infraestrutura local ou na nuvem, no Windows ou no Linux, o Tableau aproveita seus investimentos prévios em tecnologia e se adapta às suas necessidades à medida que seu ambiente de dados evolui e cresce. Explore todo o potencial dos seus recursos mais valiosos: os dados e as pessoas.

Recursos adicionais

Avaliação gratuita do Tableau

Guia da ajuda do Tableau Online: Google BigQuery

Tableau Server e Google Cloud Platform: business intelligence de alta velocidade na nuvem (em inglês)

Soluções da Tableau e do Google

Tableau e Big Data: uma visão geral

Por que usar a análise empresarial na nuvem?

Como criar pastas de trabalho eficientes