Google BigQuery e Tableau: Práticas recomendadas

tableau

Google BigQuery e Tableau: Práticas recomendadas

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 programar 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.

Martin Sleeman, gerente de produto, Tableau
Marc Lobree, consultor de produtos, Tableau
Vaidy Krishnan, gerente sênior de marketing de produto, 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

Conteúdo

Visão geral da tecnologia

Google BigQuery

O BigQuery pode processar petabytes de dados em questão de segundos em SQL simples, sem que você precise fazer ajustes ou ter conhecimentos específicos. 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 automaticamente adiciona capacidade de processamento 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 o nosso guia na Ajuda on-line que explica como migrar do SQL herdado no site do Google Cloud Platform.

Tableau

A Tableau ajuda as pessoas a ver e a entender os dados. Nossos produtos de software colocam o poder dos dados nas mãos de pessoas comuns. Isso permite que um amplo grupo de usuários interaja com seus dados, faça perguntas, resolva problemas e agregue valor. Incorporando uma tecnologia desenvolvida na Universidade Stanford, nosso produto reduz a complexidade, a inflexibilidade e os custos associados aos aplicativos tradicionais de business intelligence. Qualquer pessoa familiarizada com o Excel pode utilizar a interface do Tableau Desktop com o recurso arrastar e soltar para criar visualizações interativas sofisticadas e painéis avançados, e compartilhá-los com segurança entre organizações usando o Tableau Server ou Tableau Online.

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 60 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.

Práticas recomendadas de desempenho: Tableau

Aproveite o Tableau 10

Uma das maneiras mais fáceis de acelerar o desempenho é assegurar que você esteja usando o Tableau 10. Manter sua implantação atualizada permite que você aproveite os benefícios de todos os aperfeiçoamentos de desempenho que adicionamos regularmente ao produto.

O Tableau 9 foi um divisor de águas para nós. Ele representou um grande salto na evolução do produto, pois incluía um número surpreendente de melhorias de desempenho inovadoras, adicionadas para garantir a responsividade das visualizações.

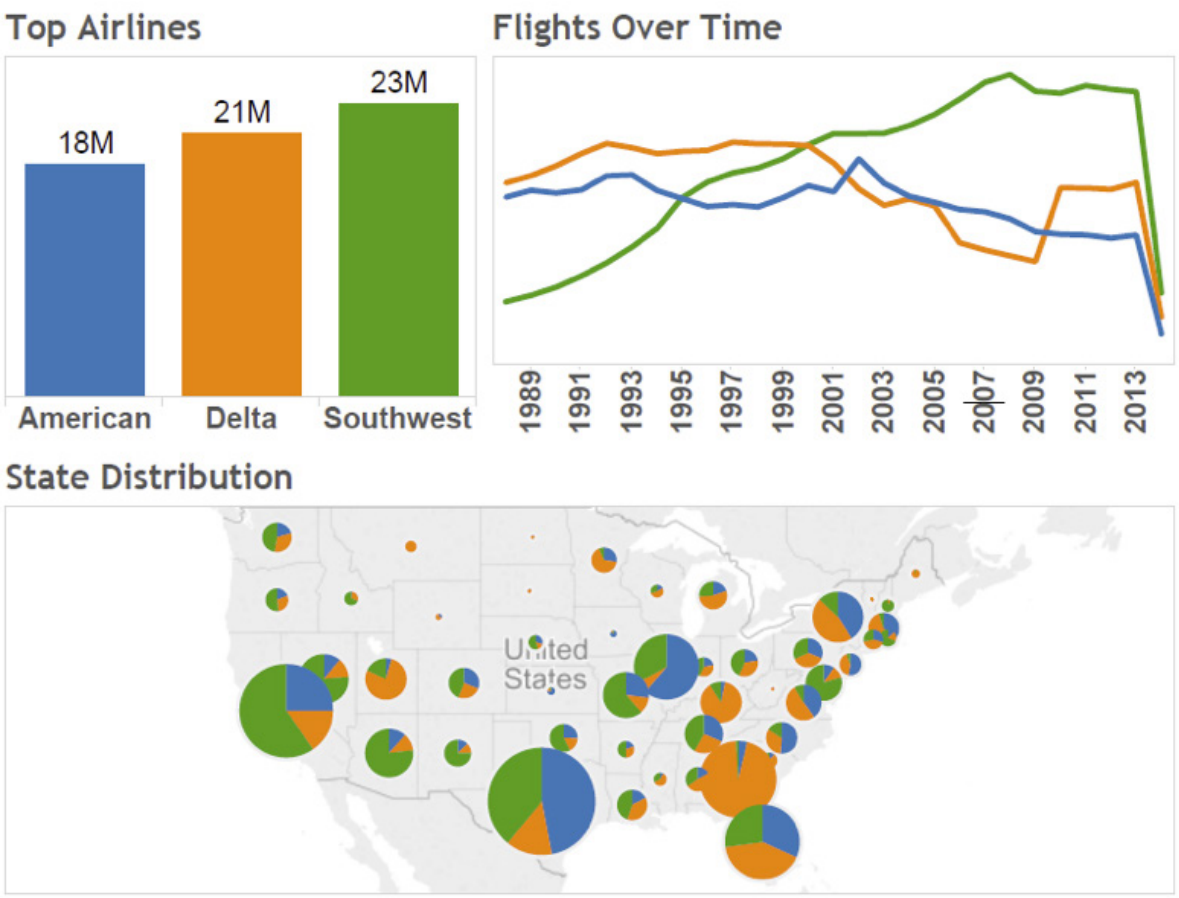
Veja abaixo algumas dessas melhorias:

- **Cálculos de nível de detalhe:** as expressões de nível de detalhe (LOD) nos permitem ir além do nível de detalhe da visualização. Os dados da visualização normalmente são o resultado da filtragem dos dados da fonte de dados. As expressões de LOD podem detectar os dados antes que eles sejam filtrados, possibilitando análises mais avançadas.

- **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 são 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 externas:** 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. Por exemplo, uma pasta de trabalho com um arquivo de extração de dados do Tableau com 157 milhões de linhas abre 50 vezes mais rapidamente quando armazenada em cache no Tableau 9 do que sem cache no Tableau 8.3.



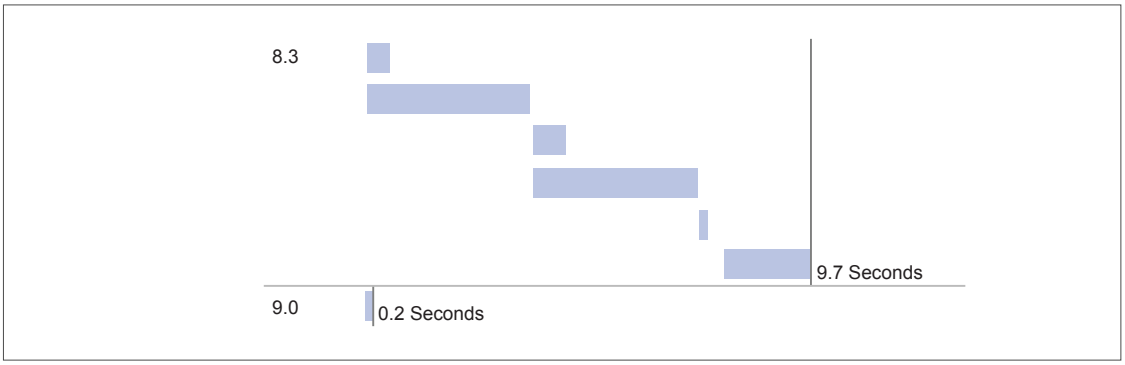


Figura 1: Melhoria de desempenho para um conjunto de dados de 157 milhões de linhas no Tableau 9.0 em comparação com o Tableau 8.3.

Com o Tableau 10, continuamos nesse caminho de inovação para garantir o desempenho de implantações empresariais escalonáveis. Em especial, adicionamos:

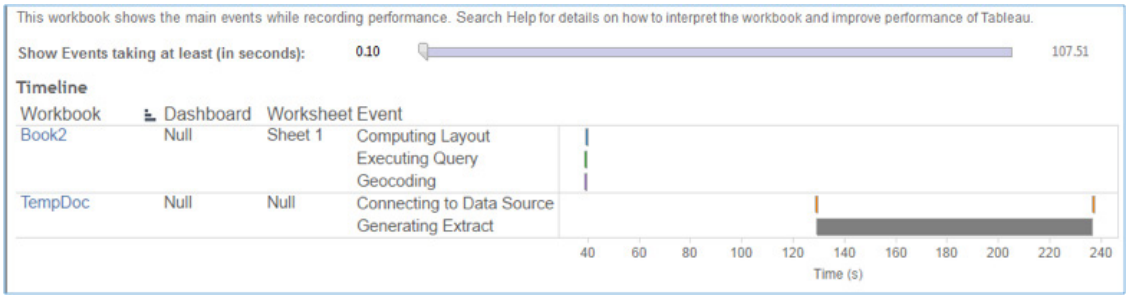
- **Desempenho aprimorado no navegador:** no Tableau 10, colocamos o carregamento da pasta de trabalho inicial em cache para agilizar o desempenho da pasta de trabalho. Aproveite tempos de carregamento mais acelerados, atualizações instantâneas baseadas em suas interações e muito mais.

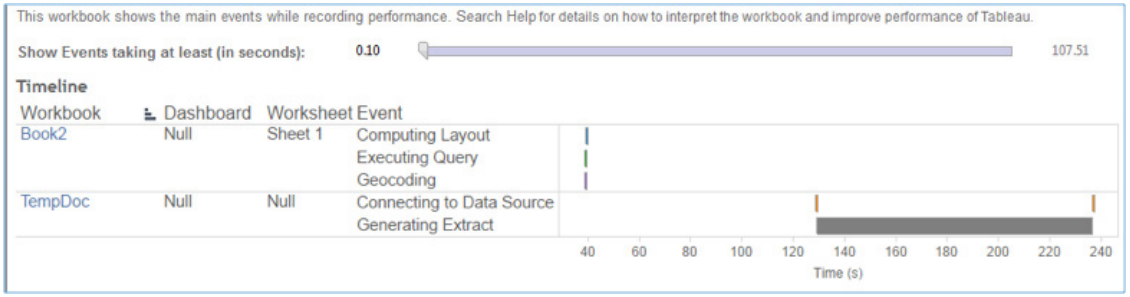
- **Conexões sob demanda no Tableau Desktop:** quando você abre uma pasta de trabalho publicada, o Tableau Desktop apenas se conecta às fontes de dados necessárias para exibir os dados da planilha atual. Em outras palavras, veja seus dados com muito mais velocidade.

- **Melhorias na estabilidade do Tableau Server:** o Tableau 10 trouxe muitas melhorias para a estabilidade do Tableau Server. Para as instalações em um único servidor, tornamos o Tableau Server mais robusto para lidar com problemas durante períodos com latência alta de entrada e saída no disco. O Tableau Server exige três nós para garantir a alta disponibilidade, mas, anteriormente, permitimos que os clientes configurassem o failover e a replicação com dois nós, o que os deixava confusos. Agora, em instalações com dois nós, apenas uma única instância do repositório será permitida. Se você precisar de failover ou alta disponibilidade com uma segunda instância do repositório, instale o Tableau Server em um cluster com pelo menos três nós. Dessa forma, você pode configurar duas instâncias do repositório e se beneficiar do failover automático. Também reduzimos o uso total de memória dos processos do Tableau Server para melhorar o desempenho do produto.

Registrador de desempenho

O Registrador de desempenho é uma ferramenta integrada eficiente que permite identificar consultas lentas e otimizar ao máximo o desempenho das pastas de trabalho. Ele faz isso ao monitorar o tempo que uma pasta de trabalho individual gastou para executar uma consulta e computar 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.





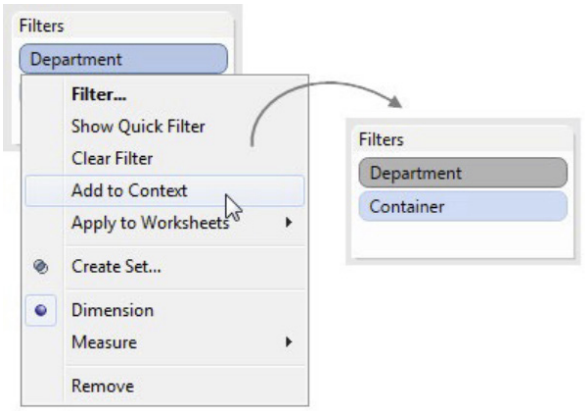
Para obter instruções sobre como criar ou interpretar registros de desempenho, consulte um destes links:

- *Registrador de desempenho no Tableau Desktop*
- *Registrador de desempenho no Tableau Server*

Filtros de contexto

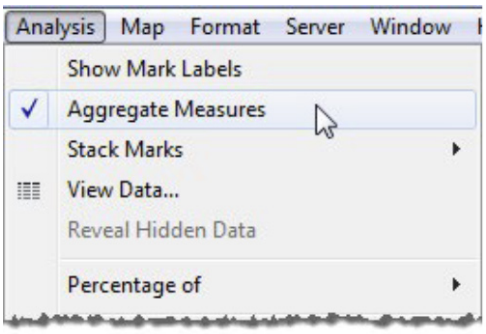
Se você for aplicar filtros a uma fonte de dados grande, poderá melhorar o desempenho configurando filtros de contexto. Um filtro de contexto é aplicado à fonte de dados primeiro, para que filtros adicionais sejam aplicados somente aos registros resultantes. Essa sequência evita que cada filtro seja 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. Consulte o nosso guia na Ajuda on-line sobre como melhorar o desempenho de exibição com filtros de contexto e aprender a criar filtros de contexto.



Agregar medidas

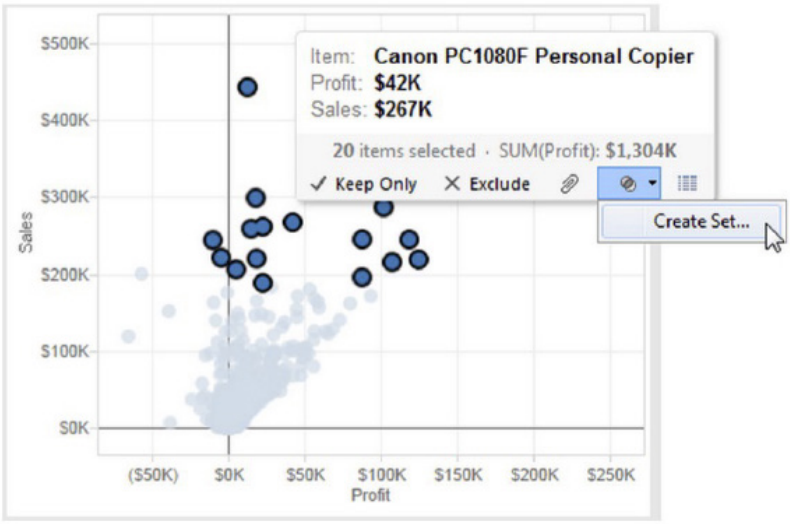
Se as exibições que você cria são lentas, verifique se está trabalhando com medidas agregadas em vez de medidas desagregadas. Quando as exibições são lentas, geralmente significa que você está tentando visualizar muitas linhas de dados de uma só vez. Você pode reduzir o número de linhas agregando os dados. Em outras palavras, verifique se a opção Agregar medidas no menu Análise está selecionada. Para obter mais informações, consulte o nosso guia na Ajuda on-line sobre como agregar dados.



Conjuntos

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. Para obter mais informações, consulte o nosso guia na Ajuda on-line sobre como criar conjuntos.

Ao criar um grupo a partir de uma seleção, como descrito em nosso guia na Ajuda on-line sobre como criar grupos, inclua somente as colunas relevantes. Cada coluna adicional incluída no conjunto resultará na queda do desempenho.



Adicionar filtros primeiro

Se você estiver trabalhando com uma fonte de dados grande e as atualizações automáticas estiverem desativadas, adicionar filtros à exibição pode criar uma consulta muito lenta. 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 degradar consideravelmente o desempenho do sistema. Nesse caso, você pode instruir o Tableau a desativar as consultas enquanto cria a exibição. Em seguida, você pode ativar novamente as consultas quando estiver pronto para ver o resultado. Consulte o nosso guia na Ajuda on-line sobre como gerenciar consultas para obter mais informações.

Atenção aos 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.

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 dá suporte a 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.

Como o armazenamento do BigQuery tem um custo muito baixo e é altamente dimensionável, geralmente é recomendável desnormalizar e pré-unir os conjuntos de dados em tabelas homogêneas. Você basicamente troca os recursos computacionais pelos recursos de armazenamento (estes últimos são mais econômicos e apresentam melhor desempenho). Como o BigQuery é um armazenamento colunar, essa troca não é uma escolha ruim, porque ele compacta melhor os dados.

O BigQuery é uma excelente ferramenta de extração, transformação e carregamento (ETL), que permite executar transformações e pipelines imensos 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, leia os documentos on-line listados abaixo.

- cloud.google.com/bigquery/preparing-data-for-bigquery#denormalizingdata
- cloud.google.com/bigquery/querying-data#largequeryresults

Segmentar tabelas por data

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 o nosso guia na Ajuda on-line sobre como se conectar a uma consulta SQL personalizada. Por exemplo, nomeie suas tabelas da seguinte forma:

mytable_20170501, mytable_20170502, etc.

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

```
SELECT
      nome
FROM
      `myProject.myDataSet.mytable_*'
WHERE
      idade >= 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 entender mais a fundo como trabalhar com curingas, clique aqui.

Especificar uma tabela de destino ao executar várias consultas semelhantes

Embora o cache de consultas seja útil quando executamos 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 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.

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

A Tableau ajuda as pessoas a ver e a entender os dados. O Tableau possibilita que qualquer pessoa analise, visualize e compartilhe informações rapidamente. Mais de 29.000 contas de usuário obtêm resultados rápidos com o Tableau, no escritório e em dispositivos móveis. Além disso, dezenas de milhares de pessoas usam o Tableau Public para compartilhar dados em seus blogs e sites da Web. Baixe a versão de avaliação gratuita em tableau.com/pt-br/trial e veja como o Tableau pode ajudar você.

Recursos adicionais

Baixe a versão de avaliação gratuita

Whitepapers relacionados

Por que usar a análise empresarial na nuvem?

Dez principais tendências da nuvem para 2017

Tableau Server e Google Cloud Platform: business intelligence ágil na nuvem.

Ver todos os whitepapers

Explore outros recursos

Demonstrações de produtos

Treinamento e tutoriais

Comunidade e Suporte

Histórias de clientes

Soluções