Qual ferramenta de BI para o BigQuery?

Rémy DAVID Segue
4 de abril · 6 minutos de leitura ★

Foto por criadores de campanha no Unsplash

Neste artigo, não vou aprofundar os prós e contras das inúmeras ferramentas de BI no mercado; em vez disso, focaremos nos requisitos específicos para BI de sucesso usando o BigQuery e como as ferramentas de BI mais populares se comparam nesse contexto específico a partir de hoje (Abril de 2019).

Permissões de usuário

A primeira coisa que você esperaria do seu projeto e organização de BI é garantir a segurança dos dados corporativos, garantindo permissões de usuário adequadas. Dessa forma, os analistas e os visualizadores de relatórios só acessarão os dados que podem ver. Isso deve ser bastante trivial, dados os recursos avançados de controle de acesso do BigQuery, mas esse tópico é surpreendentemente ignorado pela maioria dos fornecedores de BI.

Vamos ver as duas principais opções relacionadas à permissão do usuário com o BigQuery.

Lidar com permissões de usuário no nível do aplicativo

Com essa opção, o aplicativo de BI se conecta ao BigQuery usando uma conta de serviço do GCP. A conta de serviço tem acesso a todos os dados no conjunto de dados subjacente e o aplicativo de BI é responsável pela aplicação de permissões de usuário e segurança de nível de linha (RLS): retornando linhas dependendo do usuário que está executando a solicitação.

Isto tem várias desvantagens:

- Todos os aplicativos devem ser sincronizados com um provedor de identidade e mapear propriedades do usuário para permissões no nível do aplicativo (não triviais) ou usar associações complexas em tabelas de direitos do usuário em todas as solicitações
- Precisamos implementar as permissões novamente para cada novo aplicativo que se conecta ao BigQuery, aumentando o risco de erros, manutenção e esforços de migração
- Os logs do BigQuery não podem ser auditados para rastrear qual usuário acessou esses dados confidenciais ou qual usuário fez essa solicitação custosa em um banco de dados Tb

E alguns benefícios:

- O cache do BigQuery pode ser usado mesmo com o RLS, já que as mesmas solicitações vindas de usuários diferentes serão idênticas (o que não é o caso usando as visualizações autorizadas do Big Query RLS)
- Às vezes precisamos dar acesso a usuários fora da organização e nem sempre queremos gerenciar esses usuários no nível do data warehouse. Nesse caso, as contas de serviço fornecem uma camada de abstração para esses usuários

Como a implementação é fácil, a maioria das ferramentas de BI e visualização de dados no mercado que afirmam oferecer suporte à conexão do BigQuery está realmente usando uma conta de serviço,

incluindo o **Looker** e o **Chartio,** mas com a notável exceção do **Power BI** , **Tableau** , **Qlik** e **Data Studio** .

Lidar com permissões de usuários no nível Big Query

Com essa opção, as permissões de usuário são configuradas no GCP / BigQuery (usando visualizações autorizadas e funções do IAM) e as credenciais dos usuários são passadas pelo aplicativo para o BigQuery para cada solicitação.

Isso tem vários benefícios:

- Os aplicativos só precisam autenticar os usuários (o que é simples usando a conexão SAML ou openID), a autorização é feita no nível do BigQuery
- As permissões de usuários são centralizadas no BigQuery e residem nos dados protegidos, garantindo que todos os aplicativos conectados ao BigQuery tenham o nível certo de segurança
- Não delegamos um componente de segurança crítico do nosso data warehouse a aplicativos de terceiros, reduzindo a superfície de ataque
- Os registros do BigQuery podem ser usados para auditar quem acessou quais dados, quando e de qual aplicativo

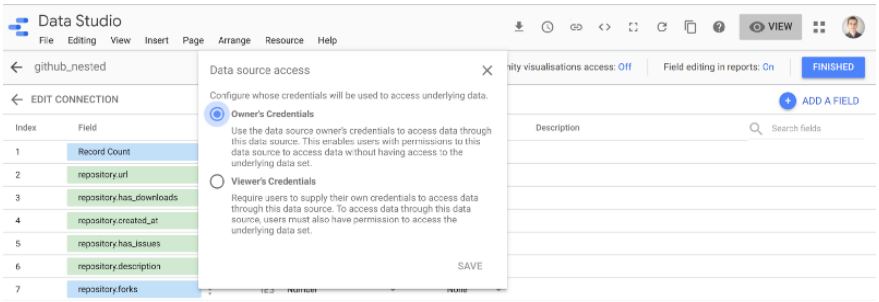
E uma desvantagem:

- O cache do BigQuery não pode ser usado, o que pode ser um problema, dependendo do número de usuários e do tráfego de aplicativos

Como a implementação dessa opção exige mais trabalho dos fornecedores para transmitir o token de usuário ao BigQuery, apenas algumas ferramentas de BI oferecem suporte a esse modelo de permissão do usuário. Dependendo da implementação, as credenciais podem ser tanto do visualizador do relatório (bom) ou do criador do relatório (ruim: menos recursos de auditoria, pode ser um problema se o criador deixar a empresa).

O Data Studio permite que você escolha entre as credenciais do visualizador e do proprietário , o **Power BI** usa as credenciais do

visualizador, enquanto o **Tableau** usa as credenciais do criador.



Opções de credenciais do proprietário do Data Studio versus visualizador em uma fonte de dados do Big Query

Para um exemplo de um caso de uso avançado usando esse modelo, veja este excelente artigo do Rich Kadel usando o Data Studio .

Faturamento

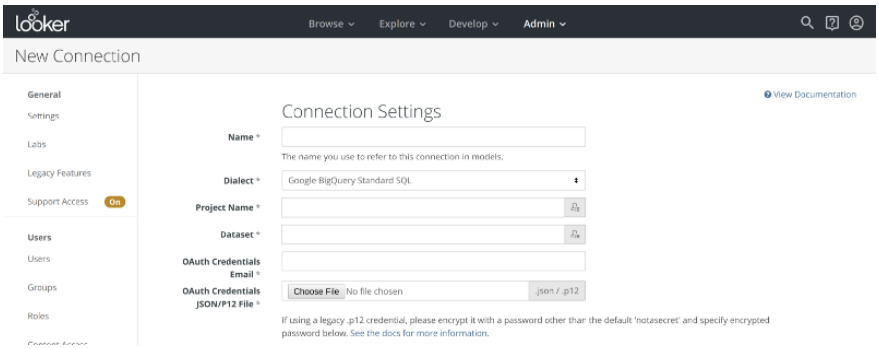
Em grandes organizações, os custos de TI costumam ser faturados do departamento de TI para as unidades de negócios. Precisamos de uma maneira de rastrear com precisão cada custo do projeto de BI. Felizmente, o faturamento de projetos do GCP faz exatamente isso: ele centraliza os custos do GCP em cada projeto.

De volta ao BigQuery, quando um usuário faz uma solicitação, há dois parâmetros por trás da cena: o *projeto da fonte de dados* e o *projeto de faturamento* . O projeto da fonte de dados é o projeto que hospeda os ***dados*** (o proprietário dos dados). O projeto de faturamento é o projeto que manipulará o ***custo*** da solicitação. Portanto, se o projeto A estiver acessando dados do projeto B, somente o projeto A será cobrado pela solicitação.

Em organizações com uma estratégia da Datalake baseada no BigQuery, por exemplo, um projeto do GCP hospedará todos os dados da empresa, enquanto os aplicativos e projetos de negócios o acessarão com seu próprio projeto de faturamento. Para as agências, um projeto hospedará dados comuns, enquanto cada cliente terá seu próprio projeto de faturamento para solicitar esses dados, etc.

Editando uma conexão de consulta grande para alterar a configuração de faturamento do projeto no Power BI

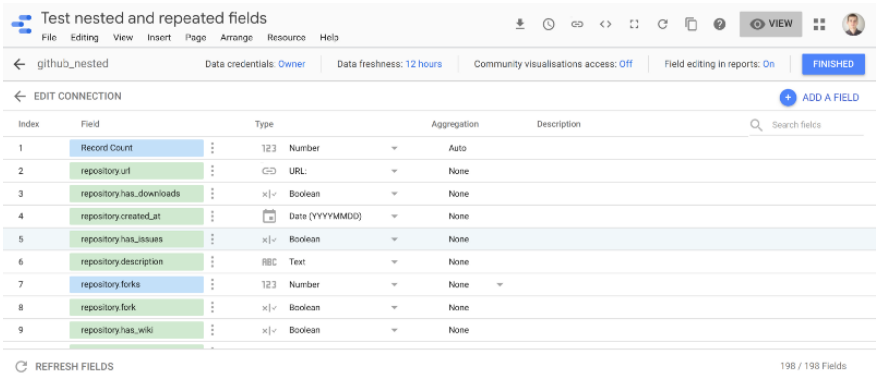
A maioria das ferramentas de BI padrão o projeto de faturamento para o projeto de fonte de dados sem uma maneira de alterar essa configuração, incluindo **Qlik** , **Chartio** e **Grow** . Com o Power BI, você precisa editar manualmente a configuração do script M da conexão, o que não é muito fácil de usar. Considerando que o **Looker** , o **Tableau** e o **Data Studio** lidam corretamente com essa configuração.



Configurações de conexão do Big Query no Looker: o arquivo de credenciais .p12 contém a configuração do projeto de faturamento

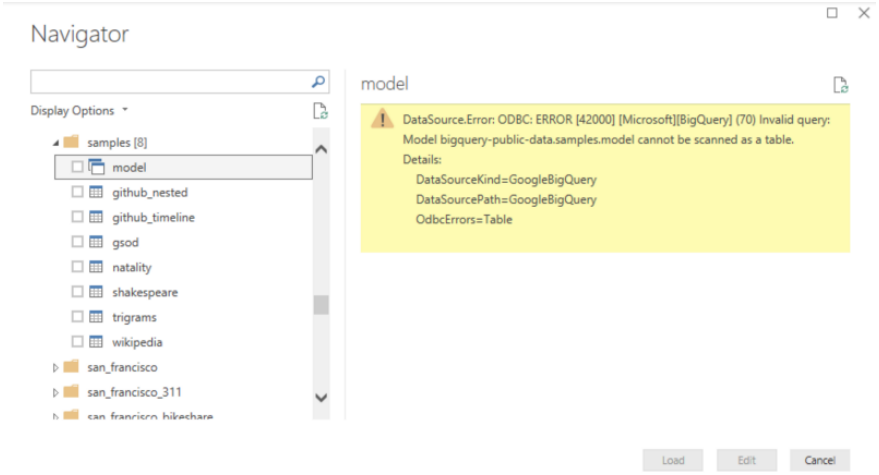
Campos aninhados e repetidos

O BigQuery tem opções muito poderosas de formatação de dados desnormalizados com campos aninhados e repetidos . Essa é a maneira mais eficiente de armazenar estruturas de dados complexas no BigQuery. Se você levar a sério o custo de armazenamento e solicitar desempenho, provavelmente já os está usando.



O Data Studio reconhece automaticamente campos aninhados e repetidos

Mais uma vez surpreendentemente, a maioria dos fornecedores de BI não suportam a exploração de tabelas usando campos aninhados e / ou repetidos em seus esquemas, exceto o **Data Studio** .



O Power BI se recusando a visualizar um conjunto de dados usando campos aninhados ou repetidos

Para as outras ferramentas, você precisará primeiro escrever uma consulta personalizada nivelando / desnegando o esquema antes que os dados possam ser explorados na ferramenta. Isso é aceitável para analistas técnicos capazes de escrever solicitações complexas do BigQuery, mas muito menos para habilitar o BI de autoatendimento para usuários casuais. Aqueles que realmente tentaram desenterrar um campo repetido em uma solicitação do BigQuery sabem o que quero dizer. Observe que isso é um problema menor com o Looker devido à sua abordagem de BI centralizada.

Tabelas federadas do Planilhas Google

Um bom recurso do BigQuery é a capacidade de analisar dados existentes do Planilhas Google no BigQuery . Assim, é possível combinar dados das fontes de dados do Planilhas Google e do BigQuery na mesma solicitação.

No entanto, para acessar uma tabela federada do Planilhas Google a partir da BigQuery API, o token oAuth da solicitação precisa incorporar o escopo do Google Drive. Algo que a maioria dos fornecedores não faz, com a notável exceção do **Data Studio** e **Metabase** (porque eu abri um problema em seu Github e eles rapidamente o implementaram).

Conclusão

Vimos que poucas ferramentas exploram todo o potencial do BigQuery para aplicativos corporativos de BI e, dependendo das suas necessidades, você precisa escolher cuidadosamente entre uma ou outra. No entanto, podemos dizer que o Data Studio tem, como seria de esperar, o melhor nível de suporte para o BigQuery, e eu definitivamente o recomendaria para BI de autoatendimento para um público amplo. O Tableau e o Looker seguem de perto e são mais direcionados a analistas / profissionais de TI e casos de uso avançados.

Obrigado pela leitura! Se você gostou deste artigo, não hesite em compartilhá-lo.

Estou aprendendo todos os dias, se eu cometer algum erro, sinta-se à vontade para me corrigir e adicionar suas sugestões na seção de comentários.