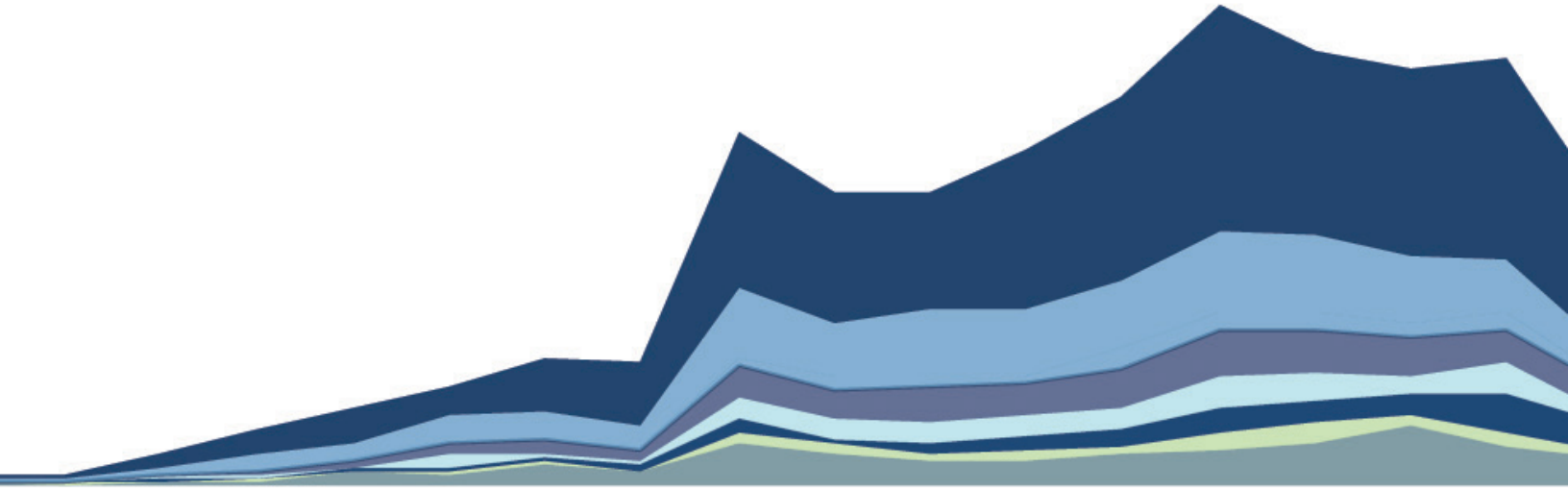


Jeff Feng, gerente de produto da Tableau

A visão da Tableau sobre Big Data

Estamos vivendo em uma nova era. Os dados são a principal “matéria prima” que orienta os negócios hoje e, com eles, faremos a próxima revolução industrial. Os novos processos industriais dos séculos XVIII e XIX transformaram completamente a forma de fabricar produtos e, de forma semelhante, a era do Big Data está redefinindo a forma de gerar, analisar e consumir dados.

A ironia é que o Big Data é tanto uma promessa quanto um risco. Os ativos de dados são cada vez mais uma área de diferenciação importante entre o sucesso e o fracasso das empresas. Contudo, a magnitude da escala, do crescimento e da variedade dos dados tornou-se tão grande, e tão dispendiosa, que sistemas de gerenciamento de bancos de dados relacionais já não são suficientes.

Com isso, as empresas estão adotando tecnologias de Big Data, como Hadoop, Spark e bancos de dados NoSQL para atender às suas crescentes demandas de dados. Elas estão usando, simultaneamente, modelos locais e na nuvem para implantar essas tecnologias. Além disso, bancos de dados de análises rápidos e data warehouses estão incorporando conceitos do Hadoop para complementar suas soluções ou diretamente criando conectores para o Hadoop. Enquanto o cenário do Big Data se desenvolve e se consolida, um aspecto persiste: as empresas precisam usar uma única ferramenta de análise para acessar seus volumes grandes ou pequenos de dados, onde quer que eles estejam.

Sumário

Estratégia de Big Data da Tableau

A Tableau tem como missão ajudar as pessoas a ver e a entender seus dados. Para cumprir essa missão, acreditamos principalmente na democratização dos dados, o que significa que “as pessoas que conhecem os dados devem ser aquelas com o poder de fazer perguntas para eles”. Os profissionais da área de conhecimento devem poder acessar facilmente os dados onde quer que eles estejam. Esses mesmos profissionais também devem poder analisar e descobrir informações sobre os dados sem a ajuda de uma minoria privilegiada: os cientistas de dados e os desenvolvedores de TI.

Independentemente do tamanho dos dados, visualizá-los é muito importante, porque eles contêm informações úteis que podem ajudar na tomada de decisão. A abordagem de visualização de Big Data é especialmente importante porque os custos com armazenamento, preparação e consulta de dados são muito altos. Por isso, as organizações devem aproveitar fontes de dados bem estruturadas e aplicar rigorosamente as práticas recomendadas para que os profissionais da área de conhecimento consultem diretamente o Big Data. Nos últimos anos, houve muitas inovações na área de Big Data. Isso resultou em uma grande diversidade de opções, cada uma com suas diferentes vantagens. A visão da Tableau ser compatível com todas as plataformas de Big Data que sejam relevantes para nossos usuários e ajudá-los a facilitar a comunicação em tempo real com seus dados.

Para concretizar essa visão para Big Data, a Tableau se concentrou em seis pilares:

1. **Acesso abrangente a plataformas de Big Data** – Parte da nossa visão é permitir a análise de Big Data, onde quer que ele esteja. O software Tableau atualmente oferece suporte a maisde 40 fontes de dados diferentes e aceita inúmeras outras por meio de nossas opções de extensibilidade. Conforme novas fontes de dados surgem e se tornam valiosas para nossos usuários, elas são incorporadas ao nosso produto para tornar mais suave o acesso aos dados.

   Nossos conectores nomeados para o ecossistema de Big Data incluem:

   - **Hadoop**: Cloudera Impala & Hive, Hortonworks Hive, MapR Hive, Amazon EMR com Impala & Hive, Pivotal HAWQ e IBM BigInsights
   - **NoSQL**: MarkLogic e Datastax
   - **Spark**: Apache Spark SQL
   - **Nuvem**: Amazon Redshift e Google BigQuery
   - **Dados operacionais**: Splunk
   - **Bancos de dados de análises rápidos**: Actian Vectorwise & ParAccel, Teradata Aster, HP Vertica, SAP Hana, SAP Sybase, Pivotal Greenplum e EXASOL EXASolution

2. **Autoatendimento para visualização de Big Data por usuários comerciais** – Os usuários comerciais podem visualizar seus dados com o recurso “arrastar e soltar”, sem precisar escrever códigos SQL/Java complexos ou trabalhos do MapReduce. O Tableau simplifica a análise de dados – Os usuários podem descobrir informações visuais em seus dados mais rápido do que nunca.

3. **Arquitetura de dados híbrida para otimizar o desempenho da consulta** – O Tableau pode se conectar em tempo real a fontes de dados ou armazená-las na memória. A conexão em tempo real é ideal para mecanismos de consulta interativos e conjuntos de dados grandes. No entanto, também podemos melhorar e acelerar fontes de dados mais lentas criando uma extração dos dados e armazenando-a em nosso Processador de dados na memória.

4. **Combinação de dados para analisar simultaneamente várias fontes de dados** – Muitas vezes, trabalhar com dados dispersos é mais complicado do que trabalhar com Big Data. É raro os dados de um analista estarem armazenados de forma organizada em um só lugar. Normalmente, eles estão espalhados e residem em tecnologias e plataformas diferentes. O Tableau permite que seus usuários trafeguem de uma fonte de dados à outra, combinando o Big Data com outras fontes de dados (por exemplo, Salesforce, MySQL e arquivos do Excel), permitindo que as organizações mantenham os ativos de dados em seus locais de origem.

5. **Desempenho global de consulta da plataforma** – À medida que os volumes de dados crescem, a Tableau continua investindo nas principais melhorias de desempenho de consultas, que ajudam a facilitar a comunicação em tempo real com os dados. Recentemente, isso inclui recursos como consultas paralelas, fusão de consultas e cache de consulta externo. Agora, o Tableau também aproveita a vetorização para processadores com suporte para esse recurso.

6. **Interface visual avançada e homogênea para os dados** – O Tableau oferece recursos de análise, como a capacidade de filtrar dados, executar previsões e fazer análises de linha de tendência com ações simples. Ele também interpreta as ações de usuário e escolhe a melhor forma de representar os dados com base nas práticas recomendadas visuais. A partir do momento que você se conecta aos dados, o Tableau também oferece uma interface visual unificada que é usada em todas as fontes de dados.

Nossa visão está alinhada com a forma como o cenário global dos dados está se desenvolvendo. O normal agora é que muitos clientes lidem com um conjunto diverso de tecnologias de Big Data. Tecnologias como Hadoop e Spark agora fazem parte da arquitetura dos dados junto com data warehouses, devido à sua capacidade de armazenar e processar dados. Em paralelo, os clientes estão ajustando o tamanho de seus data warehouses com base em suas implantações do Hadoop. Bancos de dados NoSQL são frequentemente escolhidos como back-end para aplicativos no lugar de bancos de dados relacionais, devido aos seus modelos de dados flexíveis, à sua baixa latência e ao seu design voltado para aplicativos. Além disso, as fontes de dados na nuvem também estão sempre presentes, visto que os sistemas CRM e ERP na nuvem tornaram-se a primeira opção para gerenciar processos de negócios, e o modelo de consumo pré-pago está se tornando popular para armazenamento na nuvem e processamento de dados. Com back-ends tão diversos e flexíveis, os usuários precisam de uma solução de front-end como o Tableau para se conectarem com flexibilidade a diferentes plataformas de Big Data, a fontes de dados na nuvem e a bancos de dados relacionais para terem a agilidade que desejam ao analisar dados.

Como o Tableau trabalha com Big Data?

No coração do Tableau estão o VizQL e o Processador de dados. O VizQL é uma tecnologia de propriedade da Tableau que permite ao usuário criar imediatamente uma representação visual de seus dados, fornecendo assim um feedback visual imediato. Usando o VizQL, os usuários têm uma solução de visualização unificada para produzir uma ampla gama de resumos gráficos, como gráficos de barras, gráficos de linhas e mapas com cada ação. O Processador de dados, por sua vez, é uma representação de dados em colunas, compactada e na memória, integrada à tecnologia de "conexão em tempo real" do Tableau. A tecnologia de conexão em tempo real do Tableau envia consultas SQL específicas e extremamente ajustadas da plataforma para o banco de dados, permitindo que o Tableau visualize grandes volumes de dados em tempo real, sem precisar mover os dados.

Nas próximas seções, falaremos sobre como o Tableau trabalha com Big Data para conceder acesso a dados e protegê-los, além de habilitar outros recursos especiais no Hive.

Acesso a dados

Para trabalhar com Big Data, é importante ter um modelo de conexão elegante. Nossos conectores nomeados para Big Data usam o protocolo ODBC e aproveitam os recursos específicos de banco de dados ajustando as consultas SQL enviadas:

Conexões baseadas em SQL

O Tableau é compatível com Hadoop, bancos de dados NoSQL e Spark com a utilização de SQL. O SQL gerado pelo Tableau é padronizado para ANSI SQL-92. Usar o SQL é uma solução eficaz porque ele é extremamente compacto (uma expressão), seu código-fonte é aberto, ele é padronizado, não há dependências de biblioteca e ele é muito rico e expressivo. Com o SQL, é possível expressar, por exemplo, operações de união, funções, critérios, resumos, agrupamentos e operações aninhadas.

ODBC

O Tableau usa drivers que aproveitam o padrão de programação ODBC (Open Database Connectivity) como uma camada de tradução entre o SQL e as interfaces de dados semelhantes ao SQL, fornecidas pelas plataformas de Big Data. Para o Hadoop, isso inclui interfaces como HiveQL (Hive Query Language), Impala SQL, BigSQL e Spark SQL. Para obter o melhor desempenho possível, ajustamos de forma personalizada o SQL que geramos e incluímos agregações, filtros e outras operações SQL às plataformas de Big Data.

Interfaces NoSQL

Como o nome diz, os bancos de dados NoSQL ("Não apenas SQL") podem conter dados modelados em formatos relacionais e não relacionais. Isso também significa que eles oferecem suporte a interfaces semelhantes à interface do SQL. Atualmente, o Tableau oferece suporte a MarkLogic e DataStax Enterprise como conectores nomeados usando interfaces semelhantes à interface do SQL. Com o MarkLogic, é possível conectar o texto completo ou pesquisas complexas em dados não estruturados ou em conjuntos de dados relacionais. Com o DataStax Enterprise e o Cassandra,

oferecemos suporte à interface ODBC do Hive que usa o HiveQL para acessar o armazenamento de linhas particionado do Cassandra.

Segurança dos dados

Implementar uma solução de análise visual independente e de autoatendimento só é possível em um nível organizacional quando problemas de segurança, como autenticação e acesso a dados, encontram-se devidamente solucionados. Estamos trabalhando com várias versões para permitir um acesso seguro aos dados das fontes de dados relacionadas a Big Data.

Atualmente, oferecemos LDAP ou autenticação Kerberos para que os usuários do Tableau Desktop conectem-se com segurança aos clusters do Hive Server 2 usando Cloudera Hadoop, Hortonworks Hadoop ou MapR Hadoop. Além disso, oferecemos suporte a logon único e acesso delegado com Kerberos para conexões com o Cloudera Impala. Isso amplia o suporte anterior com Active Directory, SAML e o sistema de autenticação integrado do Tableau. Para os usuários, isso significa uma experiência mais contínua, porque os usuários conectados às suas máquinas locais não precisam fazer logon novamente no Tableau Server ou em qualquer outra fonte de dados em tempo real do Impala.
Para os administradores de TI, a compatibilidade do Tableau com o Apache Sentry garante que dados confidenciais fiquem protegidos, porque os usuários exibirão apenas os dados que têm autorização para visualizar. Com o esforço conjunto da Tableau e da Cloudera para viabilizar a delegação de usuário para o Impala, podemos garantir que os usuários conectem-se ao Impala como uma fonte de dados em tempo real por meio de uma autenticação de back-end automatizada e estável. No futuro, queremos ampliar o suporte a logon único e a acesso delegado com Kerberos para muitas outras fontes de dados.

Recursos especiais no Hadoop Hive

O Hadoop tornou-se uma tecnologia de Big Data quase onipresente. O Hadoop também expande bastante o processamento de dados, que pode ser executado na camada de armazenamento, em comparação com os bancos de dados tradicionais. Dessa forma, o Tableau apresenta diversos recursos exclusivos para conexões com o Hadoop Hive. Esses recursos incluem:

- **Processamento de XML** – O Tableau oferece várias funções definidas pelo usuário (UDFs) para processar dados XML usando o XPath. Essas funções permitem que os usuários extraiam conteúdo, executem análises simples e filtrem dados XML.
- **Processamento de texto e na Web** – Além dos operadores XPath, a linguagem de consulta do Hive oferece várias formas de trabalhar com elementos comuns da Web e dados de texto, incluindo:
    - **Objetos JSON** – Recupere elementos de dados de cadeias que contêm objetos JSON.
    - **URLs** – Extraia componentes de uma URL, como o tipo de protocolo ou o nome do host, ou recupere o valor associado a uma determinada chave de consulta em uma lista de parâmetros de chave/valor.
    - **Dados de texto** – Localize e substitua texto no Hive diretamente do Tableau.

- **ETL instantâneo** – O SQL personalizado permite que os usuários definam suas conexões de dados usando condições de união complexas, pré-filtragem e pré-agregação.
- **SQL inicial** – O SQL inicial permite que o usuário especifique uma coleção de declarações SQL para execução imediatamente após uma conexão de dados ser estabelecida. Isso geralmente é feito para ajustar as características de desempenho ou desenvolver a lógica personalizada de processamento dos dados.
- **Análise personalizada com UDFs e MapReduce** – O Tableau permite que usuários implementem UDFs, funções de agregação definidas pelo usuário (UDAFs) e expressões SQL arbitrárias do Hive usando as funções de "passagem". Essas funções normalmente são criadas como arquivos Java (JAR) que podem ser copiados para o cluster do Hadoop. Os usuários também podem exercer um controle explícito sobre a execução das operações do MapReduce no SQL personalizado.

Casos de uso: Tableau e Big Data

As organizações que estão começando a usar Big Data identificarão dois casos de uso fundamentais para utilizar o Tableau e seus ativos de dados: exploração de dados e visualização de dados.

Exploração de dados

As organizações estão capturando e armazenando todos os tipos de dados, frequentemente sem predefinirem uma estratégia de análise, com a expectativa de que esses dados sejam úteis para fornecer informações no futuro. Dados como logs da Web, logs do servidor, dados de sequência de cliques, dados de sensores e dados de redes sociais agora estão sendo capturados em plataformas de dados, como o Hadoop, em vez de serem descartados. Esse tipo de abordagem oferece flexibilidade e incentiva a experimentação de diferentes tipos de análise. O Tableau pode explorar e visualizar as tendências abrangentes nos dados antes de comprometer uma quantidade significativa de recursos para transformar os dados em um produto.

Na EMC, o Tableau é usado para explorar os dados do sensor de medição de energia no Hadoop. Nas palavras do engenheiro de soluções Tom Hudgins: "tenho cerca de 70 bilhões de linhas no banco de dados que estão sendo analisadas com o Tableau. Estamos usando o Tableau para analisar dados de medição inteligente, informações sobre a energia que volta de residências e empresas. As empresa precisam saber navegar por esse fluxo de informações e extrair dele os elementos que resultarão em descobertas, fazendo-as pensar em coisas que vinham passando despercebidas; é isso que fará a diferença entre o sucesso e o fracasso."

Visualização de dados

A abordagem de uma organização para visualização de dados deve ter como foco a otimização do desempenho após a decisão de qual conjunto de análise deve ser habilitado ou transformado em produto. Expor a quantidade certa de variedades e detalhes no Big Data é essencial para alimentar

painéis responsivos e viabilizar comunicações em tempo real com os dados. Os administradores de TI e os usuários do Tableau têm recursos disponíveis para definir o escopo dos dados no nível de detalhe adequado. Além disso, também é muito importante escolher uma plataforma de processamento com suporte a análise interativa. As práticas recomendadas para maximizar o desempenho incluem:

- Aproveitar um mecanismo de consulta interativo
- Personalizar o desempenho da conexão para consultas em tempo real
- Otimizar extrações por meio de resumos, filtragens e exemplos
- Utilizar as práticas recomendadas de banco de dados, como a criação de partição

A Rosenblatt Securities é um exemplo de organização que reiterou e otimizou sua abordagem de Big Data. “Com o Tableau e uma equipe de apenas cinco pessoas, fizemos coisas que sem o Tableau e com uma equipe de até 50 pessoas levariam bastante tempo para serem feitas”, revela Scott Burrill, sócio e diretor administrativo. “Temos 800 apólices para as quais fazemos análises preditivas, a fim de determinar em tempo real se estamos em uma posição de entrada ou de saída. Com muita rapidez, conseguimos fazer análises derivadas em centenas de milhares de campos diferentes, visualizá-las, obter informações a partir delas, tomar decisões baseadas nelas e contar histórias com elas. O Tableau expandiu nossos horizontes. Agora, podemos atuar com base em informações concretas, analisando os dados como um todo, quando antes provavelmente teríamos que nos basear em amostragens.”

Resumo

A era do Big Data chegou. O volume dos dados cresce cada vez mais rápido, e as organizações estão mudando suas infraestruturas de dados para Hadoop, Spark, NoSQL e bancos de dados de análises rápidos para acompanhar o que agora é o normal em se tratando de dados. O Tableau, que é capaz de capacitar usuários comerciais comuns, como os da EMC e da Rosenblatt Securities, está levando as informações visuais do Big Data para as massas.

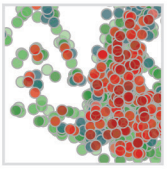
Sobre o autor

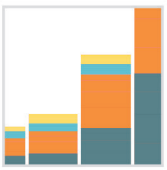
Jeff Feng, gerente de produto da Tableau Software (@jtfeng)

Jeff Feng é gerente de produto na Tableau Software e responsável pelo guia prático de produtos Big Data, estratégia do produto e desenvolvimento de novos recursos para transformar a maneira como as pessoas trabalham com dados. Antes da Tableau, Jeff foi consultor de gerenciamento na McKinsey & Co., onde prestava consultoria sobre negócios, tecnologias e estratégias de produtos para empresas de alta tecnologia da Fortune 500, bem como gerente de programa na Apple, onde ajudou no lançamento do iPhone 4. Jeff tem um MBA da MIT Sloan School of Management e um MS e BS em Engenharia Elétrica da Universidade de Illinois em Urbana-Champaign.

Sobre a Tableau

A Tableau ajuda as pessoas a ver e a entender seus dados. O Tableau possibilita que qualquer pessoa analise, visualize e compartilhe informações rapidamente. Mais de 26.000 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 www.tableau.com/pt-br/products/trial e veja como o Tableau pode ajudá-lo.





Whitepapers relacionados

Cinco práticas recomendadas para a combinação Tableau e Hadoop

Sete dicas para ter sucesso com o Big Data

Estimulando uma cultura orientada por dados:
um relatório especial da Economist Intelligence Unit em parceria com a Tableau

Big Data: a próxima revolução industrial

Tableau Software e Big Data

Grupo Aberdeen: maximizando o valor da análise e do Big Data

Consulte todos os whitepapers

Recursos adicionais

· Baixar a avaliação gratuita

· Demonstração do produto

· Treinamentos e tutoriais

· Comunidade e suporte

· Histórias de clientes

· Soluções