



Data Science Academy

www.datascienceacademy.com.br

Microsoft Power BI Para Data Science

BI e Big Data



Outra área na qual o “Self-Service” BI está decolando é a análise de Big Data. Este é um fenômeno mais recente, mas está gerando um incrível crescimento e inovação. Big Data geralmente se refere a grandes conjuntos de dados que são simplesmente muito grandes para ser gerenciados ou consultados com ferramentas tradicionais de análise de dados. Esse volume imenso de dados é resultado da explosão de ferramentas de geração de dados, rastreamento, monitoramento, transações e redes sociais (para citar alguns) que se tornaram tão populares nos últimos anos.

Não só essas ferramentas geram cargas de novos dados, mas também geram um novo tipo de dados, chamado dados "não estruturados". Em termos gerais, dados não estruturados são simplesmente dados que não foram organizados de forma predefinida. Ao contrário de dados mais tradicionais e estruturados, esse tipo de dados é composto em larga escala por texto, ao mesmo tempo que contém dados mais facilmente definidos, como datas ou números de cartão de crédito. Exemplos de aplicativos que geram esse tipo de dados incluem as ferramentas de acompanhamento do comportamento do cliente que você usa para ver o que seus clientes estão fazendo no seu site de comércio eletrônico, as pilhas de log e arquivos de eventos gerados a partir de alguns dispositivos inteligentes (como alarmes e smart sensores) e ferramentas de rastreamento de redes sociais de larga escala.

As empresas que estão implantando essas ferramentas estão sendo desafiadas não só por um dilúvio súbito de dados não estruturados que reduzem rapidamente os recursos de armazenamento, mas, ainda mais importante, estão encontrando dificuldades para consultar esta nova informação. As ferramentas tradicionais de armazenamento de dados não foram projetadas para gerenciar ou consultar dados não estruturados (Bancos de dados relacionais e Data Warehouses tradicionais). Novas inovações de



armazenamento de dados, como os Data Lakes, estão emergindo para resolver essa necessidade, mas as organizações que ainda dependem exclusivamente de ferramentas tradicionais ao implantar aplicativos que geram dados não estruturados geralmente se encontram sentadas em montanhas de dados que não sabem como usar.

Um dos principais produtos atualmente para armazenar e processar Big Data, é o Apache Hadoop, que é uma estrutura de software de código aberto que foi projetado especificamente para pesquisar grandes conjuntos de dados armazenados de forma distribuída (o que significa: no seu data center, na nuvem ou em ambos). Não só o Hadoop permite que você pesquise Big Data, ele permite que você consulte simultaneamente os dados não estruturados e estruturados. Em outras palavras, se você deseja consultar todos os dados da sua empresa para uma visão máxima, então o Hadoop é o que você precisa. O Microsoft Power BI oferece suporte ao HDInsight, o Hadoop no ambiente em nuvem da Microsoft.

Você pode baixar e implementar o próprio Hadoop para realizar suas consultas, mas geralmente é mais fácil e mais eficaz usar ferramentas de consulta comercial que empregam o Hadoop como base de pacotes de análise mais intuitivos e completos. Notavelmente, a maioria das ferramentas de Self-Service BI, incluindo Chartio, IBM Watson Analytics, Microsoft Power BI e Tableau Desktop, todos apoiam isso. No entanto, cada um exige níveis variáveis de configuração ou mesmo ferramentas de complemento para fazê-lo - com IBM, Microsoft e Tableau oferecendo recursos excepcionalmente profundos. No entanto, tanto a IBM quanto a Microsoft ainda esperam que os clientes utilizem ferramentas adicionais para a governança de dados a fim de garantir o melhor desempenho.



Se quiser aprender em mais detalhes como trabalhar com o Apache Hadoop, confira nosso treinamento totalmente prático: Engenharia de Dados com Hadoop e Spark.

Obrigado

Equipe Data Science Academy