Microsoft Power BI Para Data Science 2.0

Por que Cientistas de Dados Devem Conhecer o Hadoop?

Antes de mais nada é importante deixar claro. Hadoop não é um banco de dados. Hadoop é um framework composto de uma camada de sistemas de arquivos distribuído (HDFS) e uma camada de programação em paralelo, o MapReduce. Embora no fim das contas as duas tecnologias sirvam para armazenar e processar dados, fazem isso de formas bem diferentes e com propósitos diferentes. Vamos compreender algumas dessas diferenças. Começamos pelos bancos de dados relacionais.

Bancos de dados relacionais (RDBMS – Relational Database Management Systems) tem sido o principal modelo para a gestão de banco de dados durante as últimas décadas. Seu propósito é específico e sua presença é bastante ampla em ambientes corporativos para quase todos os sistemas informatizados.

Um banco de dados é uma aplicação que lhe permite armazenar e obter de volta dados com eficiência. O que o torna *relacional* é a maneira como os dados são armazenados e organizados no banco de dados. Quando falamos em banco de dados, aqui, nos referimos a um banco de dados relacional — RDBMS *Relational Database Management System*. Em um banco de dados relacional, todos os dados são guardados em tabelas. Estas por sua vez são um conjunto de linhas e colunas. São os relacionamentos entre as tabelas que as tornam "relacionais".

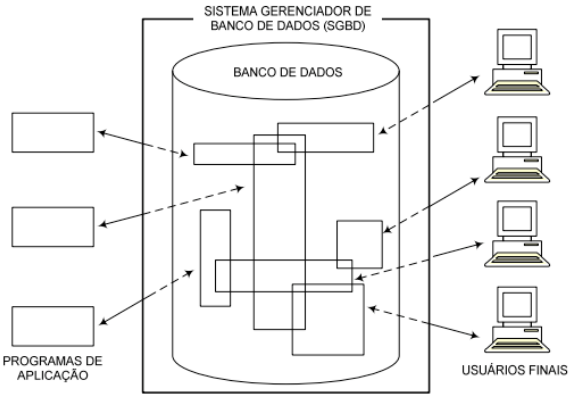


FIGURA 1 REPRESENTAÇÃO DE UM SISTEMA DE BANCO DE DADOS
DATE, 2004, P. 6 (ADAPTADO)

Durante várias décadas, os RDBMS atenderam o seu propósito e ainda o fazem muito bem, mas com o surgimento do Big Data e esta imensa quantidade de dados de diferentes categorias, gerados em diferentes velocidades, volumes e formatos, novos modelos de gestão de dados começaram a surgir. Isto levou ao crescimento por exemplo de soluções NoSQL, bancos de dados não relacionais. Com NoSQL, dados não estruturados podem ser armazenados em vários nós de processamento e não requerem schemas fixos, geralmente evitam operações de join e, normalmente, funcionam bem com escalonamento horizontal. Estima-se que existam hoje 60 bancos de dados não-relacionais e muita dessa evolução se deve ao crescimento do Big Data. O Big Data que também motivou o surgimento do Hadoop. Veja que atualmente temos diferentes opções de armazenamento para atender diferentes propósitos. Não se discute qual é melhor ou pior, mas sim qual solução deve ser usada para resolver um problema específico.

Bancos de dados relacionais usam linguagem SQL, tornando-os uma boa escolha para aplicações que envolvem a gestão de várias operações. A estrutura de um banco de dados relacional permite unir informações de diferentes tabelas através do uso de chaves estrangeiras (ou índices), que são usados para identificar exclusivamente qualquer peça atômica de dados dentro dessas tabelas. As tabelas se conectam através de chaves estrangeiras, de modo a criar uma ligação entre as suas peças de dados. Um banco de dados relacional combina dados usando características comuns encontradas no conjunto de dados e o grupo resultante é denominado como Schema.

ACID

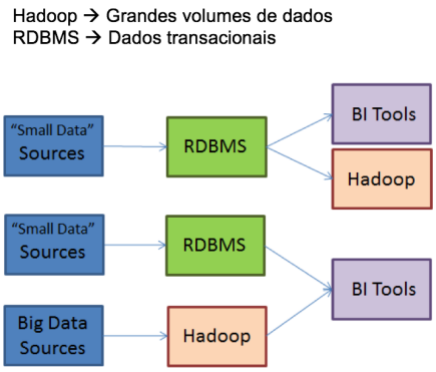
Atomicidade
Consistência
Isolamento
Durabilidade

Se você tiver aplicações que lidam com uma grande quantidade de consultas complexas, transações de banco de dados e análise de dados de rotina, você provavelmente vai preferir utilizar um banco de dados relacional. E se a sua aplicação tem como foco principal a execução de muitas transações, é importante que essas transações sejam processadas de forma confiável. Este é o lugar onde ACID (o conjunto de propriedades que garante que as transações de banco de dados sejam processadas de forma confiável) realmente importa e onde a integridade referencial entra em jogo. Principais bancos de dados relacionais do mercado:

E o Hadoop?

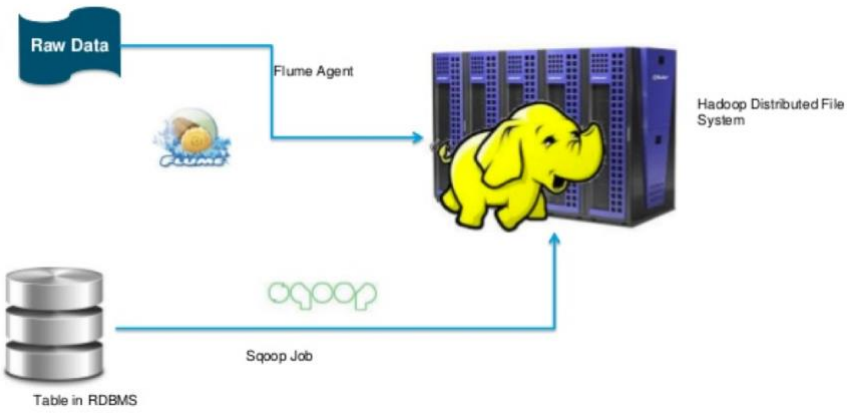
Hadoop é uma solução open-source de processamento de dados e tem como principal objetivo o processamento de dados com alto volume e variedade por meio de computação de larga escala. Com a chegada do Hadoop, o processamento massivo de dados começou a ser realizado de forma significativa.

Hadoop pode ser uma ótima alternativa para tratar, processar e agrupar grandes volumes de dados estruturados, semiestruturados e não-estruturados. Como o Hadoop pode receber dados em praticamente qualquer formato, A sua implementação oferece um meio relativamente acessível que permite extrair informações e fazer previsões a partir da compreensão dos dados da empresa, ao invés de obter informações apenas de bancos de dados transacionais ou Data Warehouses. Entretanto, alguns conjuntos de dados são realmente grandes e gerados em alta velocidade, para que o Hadoop possa tratá-los. Nestes casos, pode-se implementar soluções customizadas de MapReduce.



O Hadoop utiliza clusters para armazenar os dados, através de nodes que oferecem alta capacidade de computação quando combinados em distribuição paralela. Este tipo de solução reduz dramaticamente os custos envolvidos no armazenamento de Big Data.

O Hadoop é o local para armazenar grandes volumes de dados, normalmente não estruturados (embora ele possa armazenar dados estruturados) que podem vir de dados gerados em tempo real ou mesmo de processo em batch vindos de bancos de dados relacionais.



Hadoop processa dados em batch. Consequentemente, ele não deve ser usado para processar dados transacionais. Mas o Hadoop pode resolver muitos outros tipos de problemas relacionados ao Big Data.

Ou seja, é bem provável que Hadoop e bancos de dados relacionais convivam juntos por um bom tempo, afinal eles possuem propósitos diferentes! Agora que você compreende melhor o que é o Hadoop, você pode estar se perguntando: eu preciso aprender o Hadoop para me tornar um cientista de dados? Vejamos.

Diferentes pessoas usam diferentes ferramentas para diferentes propósitos. O termo Data Science é de certa forma um termo genérico por uma razão: um Cientista de Dados pode perfeitamente trabalhar e conduzir sua carreira sem necessariamente aprender a usar o Hadoop. O Hadoop é amplamente utilizado e embora seja a principal solução para armazenamento de Big Data, ele não é a única solução para gerenciar e manipular grandes conjuntos de dados. Você pode perfeitamente se dedicar a aprender outra solução. Mas vou dar a você 10 argumentos para ajudá-lo na decisão de aprender ou não Hadoop.

1- Hadoop é open source.
2- Hadoop oferece o framework mais completo para armazenamento e processamento de Big Data.
3- A líder mundial em bancos de dados relacionais, a Oracle, oferece soluções de Big Data Analytics com Hadoop.
4- A líder mundial em sistemas operacionais, a Microsoft, oferece soluções corporativas em nuvem, com Hadoop.
5- O Hadoop é mantido pela Apache Foundation, mas recebe contribuição de empresas como Google, Yahoo e Facebook.
6- Um Cientista de Dados deve conhecer bem o paradigma de processamento MapReduce, uma das essências do Hadoop.
7- Hadoop normalmente aparece como um dos skills mais procurados em um Cientista de Dados.
8- Por se tratar de uma tecnologia avançada, faltam profissionais de Hadoop no mercado.
9- Hadoop é usado por algumas das maiores empresas do mundo.
10- O Big Data ainda está na sua infância. Onde vamos armazenar todos esses dados?

E não é “só” isso! Um relatório de uma empresa de investimentos americana, a Avendus Capital, estima que o mercado de Big Data chegará a 60 bilhões de dólares, o que pode representar um incrível crescimento na busca por Cientistas de Dados que saibam coletar, armazenar e analisar Big Data. Mas o relatório também diz que infelizmente não haverá número de profissionais capacitados em número suficiente. Esse é um fenômeno global. Um Cientista de Dados não se forma da noite para o dia. E na verdade a coisa vai piorar. Com o passar dos anos, as vagas para Cientistas de Dados vão aumentar e o número de profissionais capacitados não vai acompanhar esta evolução.

Aprender ou não Hadoop é uma escolha sua.

Mas com certeza este conhecimento será um grande diferencial na sua carreira e na sua compreensão sobre como armazenar e analisar Big Data.