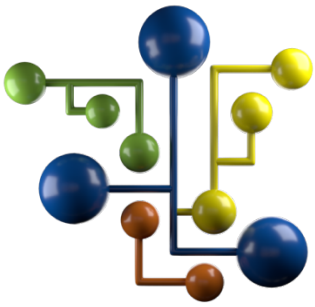


Data Science Academy

www.datascienceacademy.com.br

Engenharia de Dados com Hadoop e Spark

Apache Spark GraphX

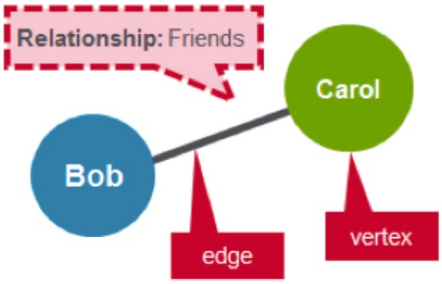
O GraphX é um dos componentes mais recentes do Spark, tendo como objetivo computação paralela de grafos.

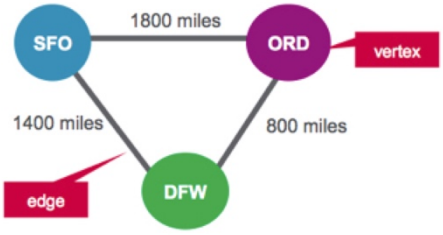
Você quis dizer grafos ou gráficos?

Boa questão. Muitas pessoas confundem esses dois termos. Vamos explicar:

grafo (Graph)
gráfico (Graphic)

Um grafo é uma estrutura matemática usada para modelar relacionamento entre objetos. Um grafo é composto de vértices e arestas. Os grafos são sem dúvida uma das formas mais interessantes de representação de informação, pois diferentemente dos gráficos, que possuem tipicamente relacionamento linear, podemos visualizar nos grafos redes interconectadas, o que representa muito melhor a relação entre objetos.





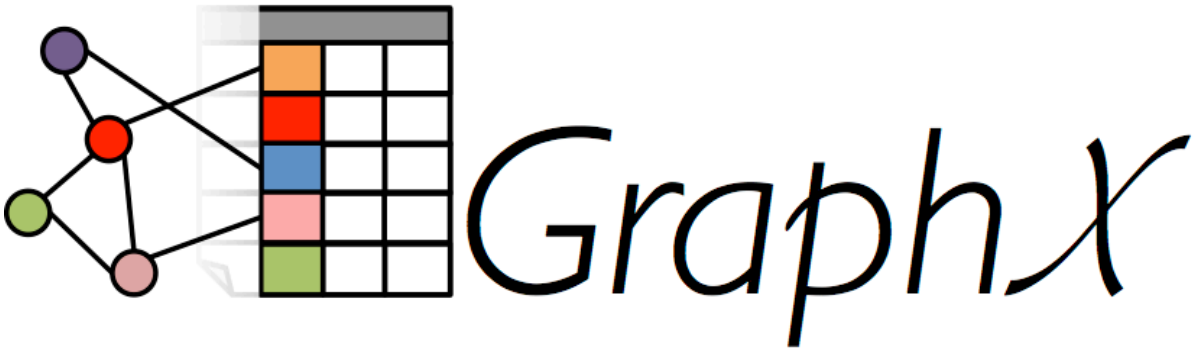
O conceito de grafos é bastante abrangente - podemos representar muitas coisas do mundo (e suas relações) utilizando esse conceito. É isso que torna a teoria dos grafos um campo tão estudado: as pessoas se debruçam para estudar esse modelo matemático porque sabem que, se conseguirem desenvolver novos trabalhos em cima desses modelos abstratos, esses trabalhos serão aplicáveis a inúmeros problemas reais. Imagine que você conseguiu desenvolver um trabalho muito bom sobre, digamos, aumentar o fluxo entre dois nós de um grafo. Esse seu trabalho poderá ser aplicado desde redes de água (aumentando o fluxo de água) até redes de computadores (aumentando o fluxo de dados transmitidos).

A teoria dos grafos é um ramo da matemática que estuda as relações entre os objetos de um determinado conjunto. Para tal são empregadas estruturas chamadas de grafos, formados por um conjunto não vazio de objetos denominados vértices e um subconjunto de pares não ordenados de vértices, chamados arestas. Para se ter uma ideia de quão importante é a Teoria dos Grafos, saiba que Google Maps e o Facebook o utilizam bastante em seus produtos. Aqui na DSA a Teoria dos Grafos é estudada em detalhes no curso Análise em Grafos Para Big Data, da Formação Inteligência Artificial.

Existem inclusive bancos de dados NoSql do tipo Graph Database, como o Neo4j e o OrientDB, além de soluções analíticas como o SAP Hana, que permite criar análises baseadas em grafos.

A análise de grafos é muito utilizada para Page Rank e filtros colaborativos, onde se busca relação entre diversos objetos. O Page Rank foi o algoritmo inicial usado nas buscas do Google.

Mas o que é o Spark GraphX? Spark GraphX é um framework para processamento de grafos de forma paralela e distribuída através de um cluster. O GraphX estende o conceito dos RDDs, criando os Resilient Distributed Property Graphs.



Ou seja, um RDD para processamento dos elementos de um grafo, como o vértice ou as arestas. Há diversas maneiras de armazenarmos grafos em computadores. A estrutura de dados usada dependerá tanto da estrutura do grafo quanto do algoritmo usado para manipulá-lo. Teoricamente, podemos dividir entre estruturas do tipo lista e do tipo matriz, mas em aplicações reais, a melhor estrutura é uma combinação de ambas.

O GraphX, por ser um componente novo do ecossistema Spark, está disponível apenas em linguagem Scala por enquanto. Criar uma aplicação analítica com o GraphX, passa pelo aprendizado da teoria dos grafos e computação paralela de grafos, elementos normalmente utilizados em Inteligência Artificial.