Tarefa Final

https://github.com/lzanotto/hmr2020

Nome

Ivan Prado da Costa

O que é GPU e para que é usada.

RESPOSTA:

A sigla GPU significa Graphics Processing Unit (ou Unidade de Processamento Gráfico) e se trata de um tipo específico de processador que atua com eficiência em paralelo e é usado para trabalhar em tarefas dedicadas, como por exemplo manipular e processar gráficos (em placas de vídeo).

Como o Spark Utiliza a GPU e mostre sua arquitetura.

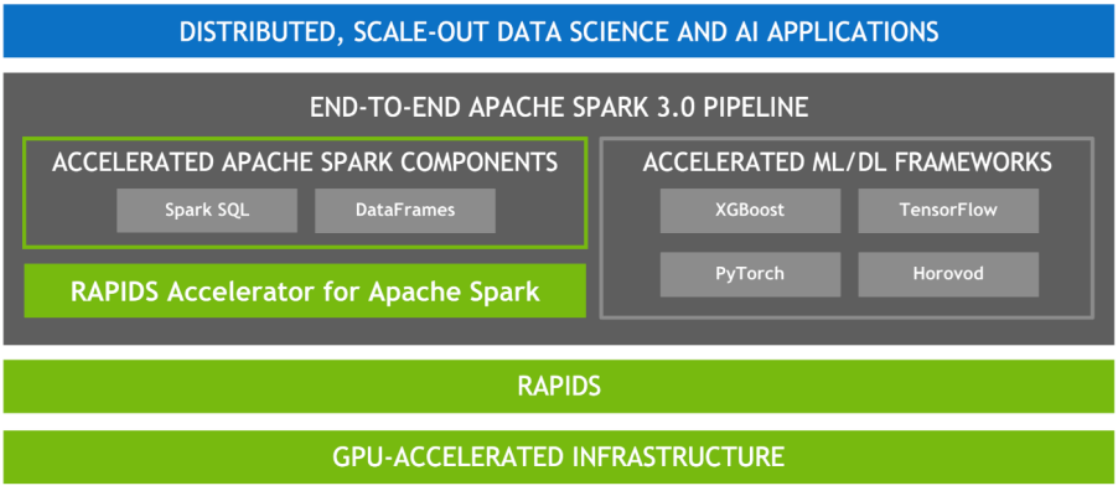
RESPOSTA:

Segundo material desenvolvido por Carol McDonald com contribuições da NVIDIA, o Apache Spark por si só já foi desenvolvido com o objetivo de manter os benefícios do MapReduce no sentido de ter uma estrutura de processamento escalável, distribuída e tolerante a falhas. Além disso, o Spark tornou esse processo mais eficiente e fácil de usar para pipelines de dados e algoritmos iterativos porque armazena dados em cache na memória e usa threads mais leves, além de fornecer um modelo de programação funcional mais rico.

O Spark atenuou os problemas de entrada/saída presentes no Hadoop, mas o gargalo mudou para processamento de um número crescente de aplicações. Isso foi tratado com o uso de GPUs.

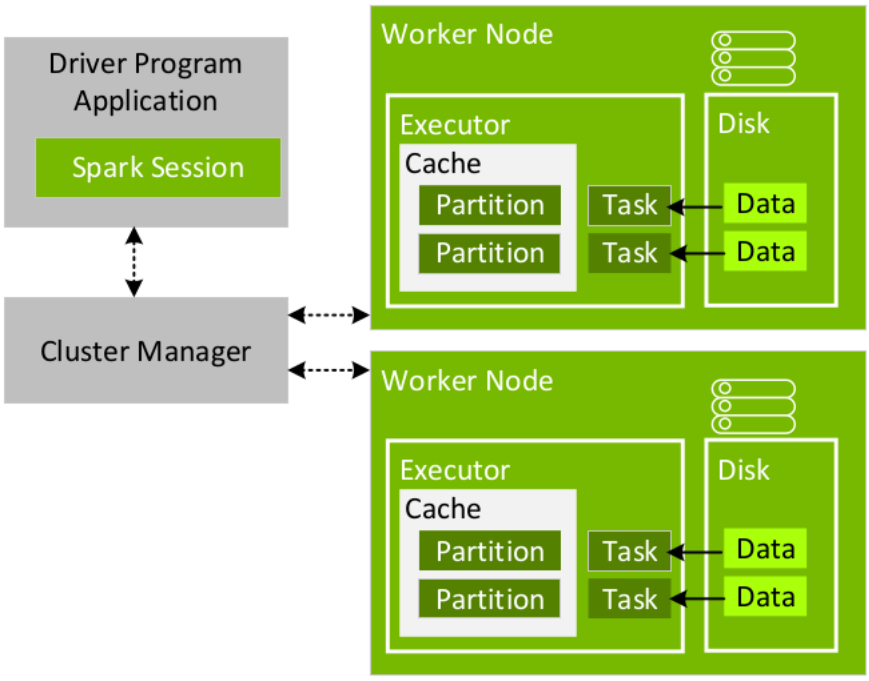
O Spark em suas versões 3.x (a partir de 3.0) adicionou funcionalidades que permitem maior integração com gerenciadores de cluster para solicitar GPUs, o que torna o processo mais fácil e rápido para o usuário.

//Imagem com estrutura de aplicações, Spark, RAPIDs e GPUs



A execução é feita em paralelo com processos em nós workers que realizam as tarefas distribuídas pelo Driver Program. Além disso, no caso de execução de forma distribuída, entra em ação o Cluster Manager que administra máquinas/clusters de acordo com as tarefas.

//Imagem com estrutura Spark em Cluster com GPU



O que é Stream e como o Spark utiliza o Stream.

RESPOSTA:

O Stream, assim como o Batch, dizem respeito a modos diferentes de entrada de dados para processamento. No caso do Batch, os dados dão entrada agrupados em lotes. Já em Stream os dados são processados de forma contínua (quase) em tempo real. O nome é também utilizado em nosso dia-a-dia para sistemas de transmissão de conteúdo (áudio e vídeo, por exemplo) via internet. É o caso de plataformas como Netflix, que realizam o chamado streaming audiovisual.

Para o Spark, é possível utilizar biblioteca e recursos adicionais, inclusive processamento por Stream para agregar em sua utilização em casos que são necessários. Como mencionado acima, é um recurso interessante para situações que precisamos que os dados sejam recebidos quase que em tempo real para processamento e análise. Exemplo atual: base de dados de apuração de votos que é atualizada de tempos em tempos pelas cidades e estados espalhados pelo país. Não é interessante esperar o final do dia para a análise. Por isso se faz necessário um recurso como entrada de dados via streaming.

O que é Machine Learning cite como o Spark utiliza sua biblioteca Mlib para acelerar o processamento.

RESPOSTA:

Machine Learning ou aprendizado de máquina é um método de análise de dados que automatiza a construção de modelos analíticos e é baseado na ideia de que os sistemas podem aprender com dados, identificando padrões e suportando decisões com o mínimo de intervenções humanas.

O Spark por sua vez também explora recursos de Machine Learning através da biblioteca MLlib composta por algoritmos e utilitários específicos, como recursos para realização de classificação, regressão linear, regressão logística, teste de mínimos quadrados, árvores de decisão, entre outros recursos contruídos com base estatística e computacional bastante robusta para a resolução de problemas.