Tarefa Final

[https://github.com/lzanotto/hmr2020](https://www.google.com/url?q=https://github.com/lzanotto/hmr2020&sa=D&ust=1603541672879000&usg=AFQjCNFp9Frcvv3Z5Zyo541dS0gvtKx55A)

1. O que é GPU e para que é usada.

GPU é uma placa gráfica chamada Device, um tipo de processador assim como a CPU, mas com muito mais vantagens, responsável por calcular e realizar tarefas determinadas pelo usuário, realizando o processamento em paralelo. No exemplo: Tesla V100. A utilização de GPU deixa o analista de dados menos tempo ocioso pois ela processa mais rápido os dados. E o ideal é ocupar a memória de cada GPU com o máximo de memória que ela tiver durante o processamento.

1. Como o Spark Utiliza a GPU e mostre sua arquitetura.

A partir da versão Spark 3.0 já existe suporte para GPU, antes disso não tem. O Spark reconhece as GPUs juntamente com a CPU e a memória do sistema. Isso permite que o Spark coloque cargas de trabalho aceleradas por GPU diretamente em servidores que contêm os recursos GPU requisitados, conforme são necessários para concluir um trabalho. Na sua arquitetura, pode ter um único pipeline, da entrada de dados à preparação de dados e ao treinamento de modelos. As operações de preparação de dados agora são aceleradas por GPU e a infraestrutura de ciência de dados é consolidada e simplificada.

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

O Stream é uma chamada API que possibilita transmitir e acessar conteúdos pela internet e em tempo real. O Spark possui dentro dele biblioteca streaming. Spark Streaming provê uma API para manipulação de dados através das RDD’s, permitindo que se desenvolva projetos de manipulação de dados armazenados em memória, disco ou sendo coletado em tempo real. Usando a API de processamento de streams, conseguimos filtrar, contar, agregar dados, fazer projeções e agrupar os dados que estão passando em tempo real em seu stream.

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

Machine Learning é um sistema que pode modificar seu comportamento autonomamente tendo como base a sua própria experiência. É um aprendizado automático que explora o estudo e construção de algoritmos que podem aprender de seus erros e fazer previsões sobre dados. Tais algoritmos operam construindo um modelo a partir de inputs amostrais a fim de fazer previsões ou decisões guiadas pelos dados ao invés de simplesmente seguindo inflexíveis e estáticas instruções programadas.

A biblioteca MLlib do Spark possui diversos tipos de algoritmos de Machine Learning, incluindo classificação, regressão, clustering e filtros colaborativos, assim como permite avaliação de modelos e importação de dados. Tudo isso foi criado para trabalhar muito bem através de cluster de servidores. Pode haver muitas etapas necessárias para processar e aprender com dados, exigindo uma sequência de algoritmos. Os pipelines definem os estágios e a ordenação de um processo de Machine Learning. No MLlib, os estágios de um pipeline são representados por uma sequência específica de PipelineStages, onde um Transformador e um Estimador executam cada uma das tarefas.

1. Reproduza um código da Web utilizando SQL com Python (pyspark). Submeta o código no github. Deixe seus comentários sobre o desenvolvimento aqui.
2. Reproduza um código da Web utilizando Stream com Python (pyspark). Submeta o código no github. Deixe seus comentários sobre o desenvolvimento aqui.
3. Reproduza um código da Web utilizando Machine Learning com Python (pyspark). Submeta o código no github. Deixe seus comentários sobre o desenvolvimento aqui. faca um gráfico mostrando a analise do resultado, envie junto no github.