Materiais:

Q: Onde encontro os materiais de apoio e os arquivos notebooks para execução dos exercícios?

R: O material de apoio consta no ícone de folha a4 ao lado do título da disciplina, os notebooks do professor Takeshi e do professor Ferreto estão divididos em suas respectivas aulas.

Na aula 01 o professor Takeshi utiliza a imagem do Hadoop em Docker, dessa maneira ele disponibilizou o GitHub que contém essa imagem. As apresentações das aulas 01 e 02 também estão nesse mesmo ícone.

Durante as suas aulas, o professor Ferreto utilizou também o Docker para apresentar os conceitos de Infraestrutura como Hadoop, Spark, Pig etc. Para pleno funcionamento do ambiente, elaboramos um tutorial para configuração do ambiente que vai desde o Docker até a instalação da imagem do Linux necessária. O tutorial se encontra no ícone a4 da disciplina, na aula 03.

Na aula 03 também tem o link do Github para download dos notebooks para as aulas práticas do professor Ferreto.

Ferramentas:

Q: Foi comentado em aula sobre o arquivo docker-compose.yml, não encontrei o memo no docker hub, onde posso encontrá-lo?

R: O docker-compose.yml está disponível no repositório https://github.com/big-data-europe/docker-hadoop.

Q: Sempre consegui realizar os exercícios propostos no curso usando o Jupyter notebook sem instalação na minha máquina, usando o Google Colab ou Kaggle. No entanto, não consigo executar os notebooks das aulas dessa disciplina. Pode ser realizado nesses ambientes ou preciso de um computador com Linux?

R: Os notebooks da disciplina foram criados para demonstrar o provisionamento do ambiente Hadoop na máquina local. Portanto, é necessário executá-los localmente. Eles não irão funcionar usando o Colab. Eles funcionarão no Linux, porém não é obrigatório um computador com o sistema operacional Linux. É necessário ter o Docker instalado e configurado na máquina local para executar os notebooks. Para isso, é possível seguir o tutorial desenvolvido pelo tutor e professor para instalar o Docker e a imagem do Linux (para Windows) e configurá-los.

Q: Tive problema ao executar o primeiro código: %load_ext dockermagic. Alguma dica?

R: Você deve baixar todo o repositório (https://github.com/tiagoferreto/HadoopJupyter). O dockermagic é um módulo para facilitar o uso do Docker dentro do Jupyter. Ele está implementado no arquivo dockermagic.py que fica juntamente com os notebooks. O comando %load_ext carrega o módulo e permite o seu uso no notebook.

Q: Eu não estou encontrando a aba Dockerfile, na página da imagem bde2020/hive. Aparecem somente as abas Overview e Tags. Gostaria de saber se tem outro caminho para encontrar este arquivo, pois não consigo dar seguimento ao exercício do professor

Takeshi. Quando tento executar o próximo comando (Docker build -t "hive:hive".), o terminal retorna uma mensagem de que não há diretório.

R: Você pode encontrar o Dockerfile e demais arquivos no repositório do GitHub (https://github.com/big-data-europe/docker-hive). Basta baixar o repositório usando o botão Code - Download ZIP ou com o comando git clone. Entrar na pasta local com o repositório (se baixar o zip, será necessário descompactar o arquivo antes) e executar os comandos para execução do container com o hive.

Q: Não entendi como e onde devo criar o arquivo 'dockerfile'.

R: O repositório https://github.com/big-data-europe/docker-hive possui todos os arquivos necessários para executar o ambiente com o Hive. Basta baixar o repositório (Code - Download ZIP ou através do git clone), entrar na pasta local com o repositório e executar os comandos para iniciar o ambiente.

Conceitos e Exercícios:

Q: O conceito central do Hadoop, plataforma para Big Data, é a localidade dos dados?

R: Processamento, de fato, também é importante no Hadoop (Map-Reduce como mecanismo, concorrência/paralelismo na manipulação dos dados). Mas mesmo estes elementos são consequência da localidade. Esta discussão ocorre na aula do Prof. Tiago Ferreto, na parte 2, aos 21 minutos.

Q: No vídeo da parte 2 da aula 03, minuto 15:15 foi abordado o conceito do shared nothing, em que os nós não compartilham informações e processam de forma independente. Fiquei com um dúvida: se nenhuma informação é compartilhada, como que é implementado o mecanismo de tolerância a falhas e retomada do processamento por outro nó caso um nó venha a cair? Não consegui estabelecer essa correlação.

R: No Hadoop cada nó processa de forma independente as tarefas que recebe do processo mestre, já que os dados (blocos) estão disponíveis localmente no nó. Existem mecanismos que permitem o mestre verificar se o nó está ativo (heartbeats). Caso seja identificado algum problema (por exemplo, queda do nó, NodeManager ou erro de processamento), a tarefa que estava sendo processada por esse nó é enviada pelo mestre para outro nó ativo que possua uma réplica dos dados do nó original.

**FAQ gerado com base em comentários até o dia 02/03/2022.*