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, no entanto a página sofreu algumas alterações e por tal motivo, criamos um tutorial para os alunos para fazer a instalação do ambiente para as aulas 01 e 02 do professor Takeshi. 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 outro 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: Estou tentando seguir os passos para a instalação do Hive no docker, porém não há aba Dockerfile na pagina, houve alguma alteração?
https://hub.docker.com/r/bde2020/hive

R: O site do DockerHub realmente passou por algumas modificações, assim elaboramos um tutorial para instalar o ambiente para as práticas do professor Takeshi, o tutorial encontra-se no ícone A4 da disciplina.

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.

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.

Q: Com base no meu hardware disponivel, como escolher o numero de executor-cores, num-executors e memory?

R: Podemos pensar em um caso genérico. Hardware : 6 nós e cada nó 16 núcleos, 64 GB de RAMCada executor é uma instância JVM. Para que possamos ter vários executores em um único NodePrimeiro: 1 núcleo e 1 GB são necessários para SO e Hadoop Daemons, portanto, estão disponíveis 15 núcleos, 63 GB de RAM para cada nóComece com como escolher o número de núcleos:Número de núcleos = tarefas simultâneas conforme o executor pode ser executadoAssim, podemos pensar que mais tarefas simultâneas para cada executor darão melhor desempenho. Mas pesquisas mostram que qualquer aplicativo com mais de 5 tarefas simultâneas, levaria a um mau show. Então coloque isso em 5.Esse número veio da capacidade do executor e não de quantos núcleos um sistema possui. Então o número 5 permanece o mesmo mesmo se você tiver núcleos duplos (32) na CPU.Número de executores:Voltando à próxima etapa, com 5 como núcleos por executor e 15 como total de núcleos disponíveis em um Node (CPU) - chegamos a 3 executores por nó.Assim, com 6 nós e 3 executores por nó - obtemos 18 executores. De 18, precisamos de 1 executor (processo java) para AM em YARN, obtemos 17 executoresEste 17 é o número que damos ao spark usando --num-executors durante a execução do comando shell spark-submitMemória para cada executor:Da etapa acima, temos 3 executores por nó. E a RAM disponível é de 63 GBPortanto, a memória para cada executor é 63/3 = 21 GB.No entanto, uma pequena memória de sobrecarga também é necessária para determinar a solicitação de memória completa ao YARN para cada executor.A fórmula para essa sobrecarga é max(384, .07 * spark.executor.memory)Calculando essa sobrecarga - 0,07 * 21 (Aqui 21 é calculado como acima de 63/3) = 1,47Desde 1,47 GB > 384 MB, a sobrecarga é de 1,47. Pegue o acima de cada 21 acima => 21 - 1,47 ~ 19 GBEntão memória do executor : 19 GB. Números finais : Executores - 17, Cores 5, Memória do Executor - 19 GB.

Q: Após solicitar a execução do comando "docker exec -t -u hadoop hadoopimg bash", estou recebendo a seguinte exceção no terminal: "unable to finda user hadoop: no matching entries in passwd file".

R: Provavelmente houve algum erro anterior na construção da imagem. A quarta célula com comandos do notebook Jupyter "1.HadoopDocker.ipynb" realiza a criação do usuário Hadoop (useradd -m -U -s /bin/bash hadoop). Além disso, o comando indicado (execução do bash dentro da imagem) não faz parte dos comandos indicados nos notebooks usados nas demonstrações. Sugiro revisar os tutoriais disponibilizados para instalação do ambiente corretamente.

** FAQ gerado com base em comentários até o dia 30/11/2022.*