UnB/FCTE – Engenharia de Software
PSPD - Programação para Sistemas Paralelos e Distribuídos Prof. Fernando W. Cruz

Atividade extra-classe – Laboratório Hadoop

A) Descrição e objetivos

Atualmente o conceito Big Data tem estado em evidência pela comunidade acadêmica porque se propõe a lidar com grandes e variados volumes de dados, formando um ambiente propício para a construção de aplicações que favorecem a tomada de decisões pelos usuários. Nesse contexto, o Apache Hadoop (https://hadoop.apache.org/) pode ser visto como uma plataforma alinhada com os requisitos do Big Data, além de ser aberta e de uso facilitado. O objetivo desse laboratório é criar um contexto para que o aluno possa *(i)* experimentar instalar, configurar, monitorar essa plataforma, e *(ii)* identificar o nível de tolerância a falhas/faltas do Hadoop e os efeitos na performance de aplicações.

B) Atividades de laboratório

Para alcançar os objetivos propostos, os alunos devem realizar as seguintes atividades:

- **Montar um cluster Hadoop básico (configuração básica)** – Os alunos devem fazer uma instalação Hadoop em modo cluster, composto por um nó mestre e pelo menos dois nós escravos (*workers*), incluindo interface web de monitoramento do *framework* e dos serviços submetidos no ambiente. Neste caso, é importante anotar e documentar todos os arquivos de configuração utilizados nos nós *master* e *slaves*, incluindo eventuais peculiaridades (diferenciais) sobre a instalação feita.
- **Teste do *framework* Hadoop** – A partir da configuração básica, os alunos devem promover algumas alterações no cluster (pelo menos 5 mudanças), de modo a gerar algum impacto no escalonador de processos (Yarn), no sistema de arquivos HDFS e no funcionamento geral de aplicações. Essas alterações podem ser feitas nos arquivos de configuração do Hadoop e os efeitos são relativos, por exemplo, ao modo como o *framework* escalona os serviços, como distribui recursos para as aplicações (memória, disco, ⋯), como o HDFS funciona, entre outros.
- **Testes de performance e de tolerância a falhas da aplicação** – Os alunos devem criar uma aplicação que leia uma grande massa de dados a ponto de a aplicação ficar em execução por um tempo razoável no *cluster* com configuração completa (*master* e todos os *slaves* ativos). A partir daí, deve-se monitorar o comportamento do Hadoop e da aplicação, considerando *(i)* o tempo de resposta (quanto demora para executar), e *(ii)* a saúde da aplicação (se mantém o funcionamento normal) sob condições adversas. Essas condições adversas devem ser provocadas em experimentos controlados e monitorados (via interface web do Hadoop). São cenários onde os nós são inseridos e retirados de formas variadas (simulando situações de falhas/faltas) enquanto a aplicação estiver em produção. Para cada condição adversa, documentar o cenário e os resultados obtidos. As conclusões sobre os testes devem estar no relatório de entrega e devem relatar se é mesmo possível melhorar o desempenho da aplicação pelo acréscimo de nós, e qual o nível de tolerância a faltas suportado pela aplicação no Hadoop. Além disso, deve-se tecer comentários sobre eventuais vantagens/desvantagens de uso desse tipo de ambiente computacional.

Obs.: Nos testes, os alunos podem variar o experimento, de modo a testar outros elementos que julgarem importantes. Por exemplo, podem envolver mais de uma aplicação wordcount (descrita a seguir) rodando ao mesmo tempo, dentre outras possibilidades.

C) Requisitos e observações sobre o experimento

Para garantir uma correta interpretação dos resultados do experimento, cabem algumas considerações:

- Como o experimento envolve análise de performance de aplicações em execução mediante inclusão/retirada de nós *workers*, as instalações com máquinas virtuais devem ser evitadas. Ainda assim, se os alunos decidirem por fazer uso de máquinas virtuais, é preciso considerar cuidados adicionais de modo a não prejudicar os testes de desempenho da aplicação em função do número de nós *workers* envolvidos. De fato, para esse laboratório, nós instanciados numa mesma máquina dificultam a percepção de alterações de performance. Em função disso, as opções de virtualização devem ser evitadas na medida do possível e o uso de *containers* não será aceito como entrega.
- A rede local do *cluster* deve ser preferencialmente com rede cabeada, se possível com uso de *switch* gigabit Ethernet. Notebooks/computadores que não tenham placas de rede podem trabalhar em rede com uso de adaptadores USB/Ethernet para realização do experimento, de modo que se possa observar a taxa de transmissão entre os computadores. Nesse caso, os grupos devem considerar a possibilidade de uso de um *sniffer* de rede para compreender o tipo e a quantidade de mensagens que flui entre os nós (por exemplo, https://www.wireshark.org/).
- Embora a especificação tenha como requisito mínimo a utilização de três nós (um *master* e dois *workers*), sugere-se a inclusão de uma quantidade maior de nós, para enriquecer os testes de carga e tolerância a falhas no ambiente.
- A aplicação a ser testada é o contador de palavras (*wordcount*) que vem como exemplo de aplicação na instalação do Hadoop. Nesse caso, basta instanciar a aplicação para leitura de um ou mais arquivos em formato texto, usando o paradigma MapReduce, desde que a massa de dados de entrada (uma biblioteca livros, por exemplo; ou um gerador de textos automático) seja suficiente para manter a aplicação executando por um tempo razoável (pelo menos 3 minutos), para que seja possível monitorá-la.
- Os resultados das execuções podem ser recuperados pelas interfaces gráficas de administração/monitoramento do Hadoop.

D) Questões de Ordem

Para este experimento valem as seguintes regras:

- O experimento pode ser feito por grupos de 4 a 5 alunos para turmas maiores de 50 alunos e 3 a 4 alunos para turmas menores; não serão aceitos trabalhos individuais. Nesse caso, basta que um dos alunos do grupo faça a postagem das entregas no Moodle da disciplina.
- A entrega é composta por *(i)* um relatório, cuja estrutura e conteúdo está descrito a seguir, *(ii)* um vídeo gravado pelos membros participantes, com apresentação do projeto. Nesse caso, considerar uma média de 4 a 6 minutos por aluno para que possam demonstar como participaram e conhecimentos adquiridos.
- O relatório a ser entregue deve conter o máximo conjunto de informações sobre o experimento (textos explicativos, figuras, roteiros de instalação, arquivos de configuração, parâmetros usados, etc.), a fim de dar qualidade ao relatório. Mais especificamente o relatório deve conter os seguintes pontos:
    - Título do laboratório, dados do curso, da disciplina/turma e identificação dos alunos participantes

- Introdução – pequena descrição da solicitação feita e uma visão geral sobre o conteúdo do relatório
- A metodologia utilizada (como cada grupo se organizou para realizar a atividade, incluindo um roteiro sobre os encontros realizados e o que ficou resolvido em cada encontro)
- Uma seção sobre a montagem do *cluster* HADOOP (básico) – apresentar um texto inicial, apresentar um diagrama da rede local montada e da disposição dos nós entre os hosts. Incluir informações sobre os arquivos de configuração utilizados nessa instalação básica e os passos para se chegar à configuração entregue
- Uma seção sobre os testes do Hadoop que eventualmente geraram alguma mudança de comportamento no *framework*
- Uma seção sobre os testes de performance e tolerância a falhas – criar cenários de teste, documentar os resultados obtidos, emitir comentários conclusivos sobre cada teste feito.
- Conclusão – iniciar com um texto conclusivo sobre o experimento e subseções para que cada aluno possa manifestar sua opinião e aprendizados específicos sobre o que foi feito, além de uma nota de auto-avaliação, em função do grau de envolvimento com o trabalho.
- Anexos (opcional) – com eventuais informações não apresentadas anteriormente, tais como arquivos de configuração, comentários sobre os códigos construídos, instruções de execução e informações adicionais para permitir replicação do laboratório pelo professor. Arquivos e informações adicionais que não puderem ser postadas no Moodle podem ser disponibilizadas via *github*.

- Caso solicitados, os alunos devem estar preparados para uma apresentação em sala, conforme definido pelo professor em data oportuna (nesse caso, trazer *slides* para facilitar a palestra).
- O laboratório será avaliado sobre dois aspectos: *(i)* qualidade das entregas, e *(ii)* participação, envolvimento com o experimento (descritos no vídeo). Com relação à qualidade das entregas, a nota é proporcional aos resultados apresentados: por exemplo, bons testes/descobertas, boa documentação e funcionalidades extras não solicitadas contam positivamente para melhorar a nota. Ou seja, os alunos podem alcançar notas melhores, quanto mais completos e bem explicados forem os experimentos. Além disso, a nota emitida levará em conta os seguintes atributos/pesos: *(i)* 25% para qualidade das entregas (relatório, vídeo, etc.), *(ii)* 15% p descrição de instalação do cluster, *(iii)* 20% p testes do cluster, *(iv)* 40% para os testes de performance e tolerância a faltas.

E) Referências

- [Tanenbaum e Steen, 2008] Tanenbaum, A. e Steen, M. V. Sistemas Distribuídos: princípios e paradigmas. 2a. Ed (livro texto do curso)
- [materiais recuperados na Internet sobre Hadoop, hdfs, map/reduce] e site oficial do Hadoop https://hadoop.apache.org/