Universidade de Brasília — Faculdade UnB Gama

Programação p Sistemas Paralelos e Distribuídos
Prof.: Fernando W. Cruz

Primeiro projeto de pesquisa - Programação de entradas *batch* em clusters

A) Objetivos do projeto

O objetivo desse projeto é permitir que o aluno avance seus conhecimentos sobre arquitetura de clusters e programação de aplicações para consumo dados em *batch* (leitura de arquivos). Esses conhecimentos devem ser adquiridos por meio da configuração de um servidor Apache Spark e um servidor Hadoop, especificamente com programação MapReduce/HDFS para classificação dos dados dos arquivos de entrada.

B) Tarefas a serem realizadas

Para atendimento a essa atividade, os alunos devem:

- Preparar um arquivo de palavras variadas, com tamanho razoável, preferencialmente acima de 800GB.
- Preparar a infraestrutura de servidores para trabalharem com programação Mapreduce/HDFS e Apache Spark, de modo que consigam ler o arquivo de entrada
- Preparar um código em versões MapReduce tradicional (https://hadoop.apache.org/) e Apache Spark (https://spark.apache.org/) para contabilizar as seguintes informações:
  a) Número total de palavras do arquivo
  b) Número total de ocorrências de cada palavra do arquivo
  c) Número de ocorrências de palavras iniciadas por cada uma das seguintes letras: S, P e R
  d) Número de ocorrências de palavras contendo cada uma das seguintes quantidades de caracteres: 6, 8 e 11.
- Anotar os resultados de processamento e desempenho das duas plataformas (Hadoop versus Apache Spark), para processamento do arquivo de entrada.
- Promover alterações nas configurações dos servidores (Hadoop e Spark), de modo a trabalharem com mais de um nó de processamento e gerar testes para efeito de comparação de performance entre essas duas soluções de processamento *batch*.

Obs.: A linguagem preferencial para entrega desse projeto é python

C) Questões de Ordem

- O projeto pode ser feito por grupos de até 3 alunos
- O projeto deve ter os artefatos entregues no Moodle até 07/8/2022 e apresentado ao professor no dia 08/8/22
- A entrega deve ser composta por: *(i) slides* de apresentação, *(ii)* relatório do projeto (descrito adiante) e, *(iii)* código criado, instruções de uso e todas as informações necessárias para esclarecimento e uso da aplicação feita (postados no Moodle ou disponibilizados no GitHub)
- O relatório do projeto deve ter a seguinte estrutura:
  i) Introdução - Descrever contexto associado a uma descrição do problema e uma visão geral da solução apresentada no relatório
  ii) Metodologia utilizada - Como cada grupo se organizou para realizar o atividade, incluindo um roteiro sobre os encontros realizados e o que ficou resolvido em cada encontro.
  iii) Descrição da solução - Essa seção deve ser organizada de modo a conter as seguintes informações: *(a)* detalhes da solução para geração das contabilizações solicitadas, *(b)* detalhes da configuração Hadoop, *(c)* detalhes da solução Apache Spark.

iv) Detalhamento da comparação de desempenho das duas arquiteturas utilizadas - incluir aqui, informações que permitam comparar essas plataformas com relação a facilidade de instalação e configuração, facilidade de programação e facilidade de uso em cada um dos ambientes testados.

v) Conclusão - Aqui deve constar resultados alcançados e limitações da solução final. Além disso, deve conter subseções relativas a cada membro do grupo para que possam se manifestar *(i)* sobre o projeto (aprendizado, sugestões de melhoria, comentários, etc.), *(ii)* sobre como participaram, e *(iii)* sobre autoavaliação - atribuição de a si uma nota de avaliação, em função da participação pessoal e do atendimento aos requisitos do projeto.

vi) Anexos com os arquivos que foram alterados em cada um dos ambientes, bem como scripts construídos para alcançar os objetivos desse projeto.

- A nota é individual e o valor máximo será dado ao aluno que demonstrar conhecimento sobre a solução e as plataformas usadas (preferencialmente em modo cluster) e aprendizado satisfatório com o projeto. Outros requisitos como cumprimento das datas e nível de colaboração e balanceamento de atividades entre os membros do grupo também serão consideradas para emissão da nota final.
- Além dos elementos citados no tópico anterior, a nota desse projeto será calculada em função dos seguintes itens: *(i)* atendimento aos requisitos definidos, *(ii)* qualidade do relatório, *(iii)* qualidade da apresentação oral e *(iv)* nível de participação do aluno no projeto.