Lab03 – Experimentando o *framework Hadoop* e o paradigma de programação *MapReduce*

A) Objetivo: O objetivo desse laboratório é criar um contexto para que o aluno *(i)* experimente o ambiente de programação *MapReduce*, *(ii)* entenda como funciona o sistema de arquivos HDFS, e *(iii)* teste a instalação do *framework Hadoop*.

B) Detalhes do laboratório

Para atender ao objetivo citado, os alunos devem, realizar o experimento em duas etapas. Na primeira etapa, os alunos devem instalar o *framework* Hadoop, disponibilizado no endereço https://hadoop.apache.org/releases.html e testar a solução do contador de palavras (código funcionando + roteiro para realização de testes). Obs.: Para compreensão do que está sendo feito, sugere-se dar uma olhada no TCC sobre programação MapReduce disponível no Moodle. Sugere-se, ainda, escolher um arquivo com uma grande quantidade de palavras para realização dos testes dessa etapa. Importante ainda é identificar quais os comandos básicos para que o HDFS faça o tratamento adequado do arquivo de entrada, de modo a espalhar seus blocos entre os *datanodes* (se o grupo tiver mais de um *host* disponível para trabalhar esse laboratório, o experimento ficará mais rico; no entanto, caso não seja possível, os alunos podem promover a instalação em um *host* único).

Na segunda etapa, os alunos devem construir uma pequena aplicação, considerando o paradigma *MapReduce*, que contabilize o número de amigos de um grupo de pessoas, conforme ilustrado na Figura 1.

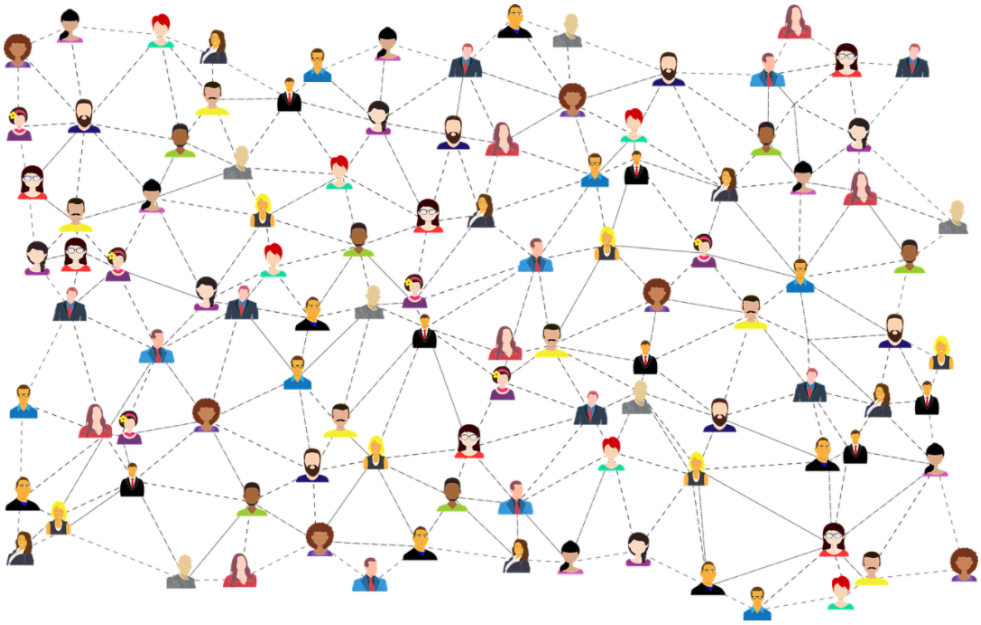

Figura 1 – Rede de amizades

Fonte: https://www.ferramentasmentais.com.br/amizades-todo-mundo-tem-mas-uma-rede-de-relacionamentos-ja-nao-e-para-todo-mundo/

Para esse laboratório, considerar um arquivo de entrada (vide Figura 2), previamente preparado, que contenha as relações de amizade presentes no grafo (considerar que cada vértice se refere a uma pessoa com nome único). Usando uma lógica *MapReduce* (similar ao contador de palavras), o programa deve gerar uma listagem contabilizando o número de amigos de cada um, conforme apresentado na Figura 1.

Arquivo de entrada	Saída esperada
Joao, Maria Maria, Joao Lorena, Jose Jose, Lorena Joao, Caio Caio, Joao Joao, Lorena Lorena, Joao	Joao 3 Jose 1 Caio 1 Lorena 2 Maria 1

Figura 2 – Formato dos arquivos de entrada e saída

Para realização dessa etapa, os alunos devem *(i)* preparar o arquivo de entrada, nos moldes do que está exemplificado, porém com dimensões grandes, a fim de testar a aplicação que foi construída, e *(ii)* construir os *scripts map* e *reduce* a serem instanciados, a fim de solucionar o que foi solicitado. Os *scripts* devem ser codificados, preferencialmente, em linguagem Java, C ou Python. Todos os resultados e conclusões com a realização dessas duas etapas devem ser documentados no relatório desse laboratório.

C) Questões de ordem

- A atividade pode ser feita por grupos de até 2 alunos. Nesse caso, basta que um dos alunos faça a postagem do material, desde que, no texto e nos códigos entregues, conste os nomes/matrículas dos membros do grupo, a fim de que sejam beneficiados com a avaliação.
- Os alunos devem estar preparados (*slides* e apresentação do código) para demonstração da solução em sala de aula, conforme sorteio definido pelo professor. O objetivo não é meramente a entrega do laboratório. Portanto, os alunos selecionados devem ser capazes de discorrer, em sala de aula, sobre a experiência com esse laboratório.
- A entrega será feita pelo envio de um arquivo zipado no ambiente Moodle da disciplina disponível em http://aprender3.unb.br. O arquivo zipado deve conter: *(i)* os arquivos com os códigos relativos às etapas 1 e 2, devidamente identificadas e *(ii)* o relatório sobre o experimento (descrito, a seguir).
- Os códigos entregues devem estar devidamente identificados, comentados e identados. O relatório do experimento deve conter os seguintes pontos:
  a) Título do experimento, dados da disciplina e do(s) aluno(s)
  b) Uma seção introdutória, com explicação sobre o *framework Hadoop*, em especial sobre o *file system* HDFS e sobre o paradigma *MapReduce*.
  c) Uma seção para cada uma das etapas do laboratório. A seção da Etapa 1 deve conter uma subseção simples descrevendo sobre o processo de instalação, e uma outra subseção apresentando os passos para realizado do contador de palavras. A seção da Etapa 2 deve conter duas subseções: *(i)* uma descrição de como o arquivo de entrada foi gerado (supondo que o grupo decidiu criar um script para geração automática do arquivo de dimensões grandes), e *(ii)* apresentação da solução para o contador de amizades.
  d) Os códigos dos *scripts* das Etapas 1 e 2 devem ser organizados em pastas no arquivo zipado, com os nomes das seções, para facilitar o teste. Nesse caso, colocar README para teste dos códigos, a fim de facilitar a correção.
  e) Opinião geral sobre esse experimento, apontando dificuldades encontradas e possíveis limitações percebidas.