UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE

CENTRO DE TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA DE COMPUTAÇÃO E AUTOMAÇÃO

**Roteiro de Aula Prática – MapReduce com Hadoop**

DISCIPLINA: DCA0132 – ENGENHARIA DE DADOS

PROFESSOR: Carlos Manuel Dias Viegas

Esta prática consiste na programação de funções map/reduce em diversos datasets para serem processados em um cluster Hadoop.

- Esta prática consiste em realizar as tarefas descritas abaixo e desenvolver os respectivos códigos fonte.

- Os códigos fonte desenvolvidos deverão ser enviados em uma tarefa específica no SIGAA.

- Esta prática pode ser realizada em duplas.

O código fonte (em Python) necessário para iniciar esta prática está disponível na seguinte página:

<https://www.dca.ufrn.br/~viegas/disciplinas/DCA0132/files/Scripts/>

**Tarefas**

Todas as tarefas abaixo devem ser executadas em um cluster Hadoop. É importante que em cada código fonte desenvolvido sejam acrescentados os comentários descrevendo o que cada etapa de código realiza.

1. Obtenção dos 10 maiores salários.
   1. Baixar o arquivo salarios.csv disponível em: <https://goo.gl/A3MhFS>
   2. Proceder às modificações necessárias nos arquivos mapper.py e reducer.py para criar as funções de mapeamento e redução com seguintes requisitos para a saída:

* Uma lista ordenada de forma decrescente com os 10 maiores salários, exibindo o nome e o respectivo salário. Deve ser considerada a penúltima coluna para obter o valor do salário.

1. Extrair nomes de arquivos com menos de 5 caracteres.
   1. Baixar o arquivo de log (weblog\_entries.txt) em: <https://goo.gl/A3MhFS>
   2. Proceder às modificações necessários nos arquivos mapper.py e reducer.py para criar as funções de mapeamento e redução com seguintes requisitos para a saída:

* Uma lista com os nomes dos arquivos .html cujo o número de caracteres seja maior ou igual a 10. A extensão .html não conta para a quantidade de caracteres.

1. Obter as estatísticas dos usuários que acessaram um serviço de e-mail.
   1. Baixar o arquivo de log do e-mail (dovecot.log) em: <https://goo.gl/A3MhFS>
   2. Proceder às modificações necessárias nos arquivos mapper.py e reducer.py para criar as funções de mapeamento e redução com seguintes requisitos para a saída:

* Uma lista em ordem decrescente com os nomes dos usuários e a quantidade de vezes que tentaram acessar o serviço de e-mail no mês de agosto, porém, só devem ser exibidos aqueles com mais de 2.000 acessos.