UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE

CENTRO DE TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA DE COMPUTAÇÃO E AUTOMAÇÃO

**Roteiro de Aula Prática – Programação pySpark**

DISCIPLINA: DCA0132 – ENGENHARIA DE DADOS

PROFESSOR: Carlos Manuel Dias Viegas

Esta prática consiste em praticar os conhecimentos de programação utilizando os conceitos de pySpark.

**Tarefas**

1. Aplicação 1: Desenvolver uma aplicação utilizando os conceitos de Spark e pySpark, que seja capaz de ler e extrair as informações do conjunto de dados disponibilizado (conjunto2.csv) a respeito dos cargos e a quantidade de pessoas que os ocupam.

Devem ser utilizadas as transformações de map() e reduce() sobre RDDs com o conjunto de dados fornecido. Pretende-se que sejam lidos todos os cargos e apresentada a quantidade de pessoas nos mesmos.

Conjunto de dados disponível em: <https://goo.gl/A3MhFS>

Datasets > conjunto2.csv

1. Aplicação 2: Desenvolver uma aplicação utilizando os conceitos de Spark e pySpark, que exiba as informações das linhas de ônibus de uma determinada cidade.

Devem ser utilizadas as transformações de map() e reduce() sobre RDDs com o conjunto de dados fornecido. Pretende-se que a aplicação seja capaz de associar as informações dos três arquivos que compõem o conjunto de dados e apresente como saída:

- O número e nome da linha de ônibus

- Os horários e o nome dos pontos em que essa linha passa

A aplicação deve aceitar como entrada do usuário o número da linha de ônibus.

Conjunto de dados disponível em: <https://goo.gl/A3MhFS>

Datasets > onibus.tar

Fonte dos dados:

Dados abertos da UFPR: http://dadosabertos.c3sl.ufpr.br/curitibaurbs/