UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE

CENTRO DE TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA DE COMPUTAÇÃO E AUTOMAÇÃO

**Roteiro de Aula Prática – MongoDB com Spark**

DISCIPLINA: DCA0132 – ENGENHARIA DE DADOS

PROFESSOR: Carlos Manuel Dias Viegas

Esta prática consiste em criar aplicações com pySpark e pyMongo que interajam com um banco de dados MongoDB por meio do Spark Streaming.

- Os arquivos com o código fonte (.py) ou (.ipynb) devem ser submetidos em uma tarefa específica no SIGAA até o dia 26/07/2024.

- Esta prática pode ser realizada em duplas.

**Tarefas**

* Desenvolver uma aplicação em pySpark, recorrendo à API pyMongo, que seja capaz de ler o arquivo de entrada abaixo a partir de uma fonte de dados (local ou HDFS) e inserir os mesmos em um banco de dados MongoDB. Pretende-se que os dados possam ser consultados (“query”) no banco de dados criado a partir da aplicação pySpark desenvolvida.
* As “queries” devem permitir realizar consultas aos dados, como, POR EXEMPLO:

- Listar todos os alunos que ingressaram por meio do SiSU

- Computar quantos alunos são do sexo masculino e do sexo feminino

- Computar o top 5 dos cursos que mais receberam alunos

- Realizar consulta múltipla ("relacional"), por exemplo: quantos alunos são do sexo masculino que ingressaram via SiSU em algum curso em específico

* Essas consultas devem passar por alguma transformação (select, filter ou outra) e o resultado dessas deve ser apresentado em tela e também escrito no banco de dados MongoDB. Recomenda-se a criação de coleções específicas de acordo com o critério de consulta para organização dos resultados no banco de dados.

PS: Não usar a biblioteca pandas. Apenas pyspark e pymongo. Utilizem dataframes.

Entrada: Lista de discentes ingressantes na UFRN em 2024

<https://dados.ufrn.br/dataset/discentes/resource/d271d842-929e-4d85-a1a8-c89cbdadb95d>

As colunas estão ordenadas da seguinte maneira:

["matricula";"ano\_ingresso";"periodo\_ingresso";"id\_curso";"id\_unidade";"id\_unidade\_gestora";"nome\_discente";"sexo";"forma\_ingresso";"tipo\_discente";"status";"sigla\_nivel\_ensino";"nivel\_ensino";"nome\_curso";"modalidade\_educacao";"nome\_unidade";"nome\_unidade\_gestora"]