Universidade Federal do Rio Grande do Norte
Departamento de Engenharia de Computação e Automação
DCA0132 – Engenharia de Dados

Trabalho prático – Projeto de Big Data

Entrega: 17/06/2018 23h59

Instruções:

- Trabalho em dupla;
- A nota deste trabalho corresponde a 80% da avaliação da Unidade II, sendo os outros 20% correspondentes à participação e assiduidade em aula;
- O trabalho consistirá na implantação de um projeto de big data, utilizando o Apache Hadoop, Spark e MongoDB;
- A submissão do trabalho deverá ocorrer via SIGAA até a data indicada acima;
- Deverá ser elaborado um relatório (sucinto, de até 10 páginas) contendo a descrição do projeto, detalhes da implantação e implementação da infraestrutura, conjunto de dados escolhido, resultados obtidos e dificuldades encontradas (caso existam);
- O trabalho deverá ser apresentado pessoalmente ao professor em horário a ser agendado.

Objetivo:

Aplicar na prática os conhecimentos adquiridos em sala de aula sobre a implantação de soluções de big data. Deverá ser criado um projeto para análise de dados que contemple a implantação da infraestrutura, bem como a análise, o tratamento dos dados e a exibição de resultados.

Forma de avaliação:

- Cada dupla irá apresentar e explicar o desenvolvimento do projeto e os resultados encontrados;
- O projeto será avaliado de acordo com a implantação e as soluções utilizadas para a obtenção dos resultados, sendo que no dia da apresentação o mesmo deverá estar operacional e ser apresentado o seu funcionamento;
- Inovações e soluções diferenciadas são bem-vindas e terão uma avaliação diferenciada (para melhor);
- Cópias não são admitidas! Evitem usar implementações de outras pessoas, sob a pena de zerar a nota.

Instruções para o projeto:

1. Implantar o Apache Hadoop (com o MapReduce e Yarn), bem como o Apache Spark e MongoDB, podendo recorrer a soluções em nuvem ou em máquina virtual;
2. Escolher um conjunto de dados (*dataset*) para execução no projeto implantado. Existem inúmeros *datasets* disponíveis na Internet, entretanto deixo como sugestão o site abaixo, no qual é apresentado um bom resumo desses conjuntos: https://paulovasconcellos.com.br/os-7-melhores-sites-para-encontrar-datasets-para-projetos-de-data-science-8a53c3b48329

a. A escolha do conjunto de dados fica totalmente a vosso critério, apenas com a restrição de que o mesmo deverá ser maior que 100 MBytes;
b. O *dataset* escolhido deverá ser tratado e inserido em um banco de dados MongoDB, e este, por sua vez, integrado ao cluster Hadoop. Pretende-se que com esta integração os dados possam ser processados e analisados para obter resultados relevantes. Como exemplos de resultados relevantes:
    i. Se considerarmos um *dataset* referente às informações climáticas dos últimos anos, a tendência se mostra de aquecimento ou resfriamento?
    ii. A taxa de sucesso dos alunos que ingressam nas universidades por meio de cotas é superior, igual ou inferior à do aluno que ingressa pelo processo seletivo (ENEM/Vestibular)?

   Portanto, o *dataset* escolhido deve permitir que sejam obtidos resultados que possam trazer conclusões relacionadas ao nosso cotidiano. Pelo menos quatro (4) resultados diferentes devem ser obtidos.

3. Para a escolha do *dataset* é necessário primeiro definir os objetivos e em seguida realizar um planejamento.
    a. Como objetivos: o que quero investigar? Que resultados posso e quero encontrar?
    b. Uma vez respondidas às questões dos objetivos, parte-se para o planejamento: O que preciso e como irei implementar a solução para alcançar os meus objetivos? Como devo estruturar e processar os meus dados? Como extrair os resultados que desejo?
4. Para inserir o *dataset* no banco de dados MongoDB, caso necessário, os dados presentes no mesmo devem ser tratados para o formato suportado pelo banco de dados;
5. Em seguida parte-se para a etapa da programação e execução:
    a. Criar uma aplicação utilizando o framework MapReduce para o processamento e extração dos resultados (4 no total);
    b. Criar uma outra aplicação utilizando o framework Spark para processamento e extração dos resultados (4 no total);
    c. Ambas as aplicações dos itens 'a' e 'b' anteriores devem, obviamente, extrair os mesmos resultados do banco de dados (ou seja, realizar as mesmas operações);
    d. Por fim, deve ser elaborada uma análise de desempenho para comparar o funcionamento, a complexidade de programação, o tempo de execução, e destacar as vantagens e desvantagens de cada um (não é desejável que se responda a este item com frases de Internet ou "achismos", mas sim baseado em resultados e testes realizados no cluster implantando).

Quesitos a serem avaliados:

- Qualidade na apresentação, organização e escrita do relatório;
- Originalidade na realização das tarefas;
- Profundidade dos detalhes abordados e soluções propostas.