PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS
COORDENAÇÃO ENSINO A DISTÂNCIA – CEAD
ESCOLA POLITÉCNICA
BIG DATA E INTELIGÊNCIA ARTIFICIAL

PROPOSTA DO PROJETO INTEGRADOR II – A

Anderson R Cunha *MSc*

GOIÂNIA,

2023

SUMÁRIO

1 OBJETIVO

Preparar e capacitar o aluno para compreender e aplicar os conceitos estudados em “Programação em Big Data (Programação em R)” e “Banco de Dados para Big Data”.

2 DESENVOLVIMENTO

Podemos definir big data de uma forma bem resumida como sendo um conjunto de conceitos e técnicas capazes de analisar grandes quantidades de dados para a geração de resultados. Essa definição apesar de bastante simplória extrapola a visão equivocada de alguns, que vem Big Data como uma ferramenta específica, como o Hadoop ou Map/Reduce, por exemplo.

Esse conjunto de conceitos e técnicas aplicados para o processamento de grandes volumes de dados não são necessariamente novos ou exclusivos do contexto de Big Data, por exemplo, muitas das tecnologias e técnicas provenientes de redes e sistemas distribuídos são aplicadas com sucesso no contexto de Big Data.

Com isso em mente, neste projeto integrador com o propósito de exercitar os conhecimentos adquiridos nos módulos de “Programação em Big Data (Programação em R)” e “Banco de Dados para Big Data”, vamos projetar um sistema de cache para otimizar o processo de plotagem de gráficos da linguagem R, adicionalmente vamos implementar na forma de protótipos, partes das rotinas desse sistema de cache com propósito de validação de conceito.

Obviamente diversas ferramenta de Big Data, incluindo bancos de dados NoSQL já fornecem caches prontos para usar mediante a configuração adequada. Entretanto nosso objetivo é reimplementar manualmente tal funcionalidade.

Requisitos do nosso sistema de cache de plotagem de gráficos:

A linguagem R fornece recursos para plotagem de gráficos, inclusive de funções pré-definidas, mediante a uma coleção de dados de entrada. Nossa necessidade consiste em: fazendo uso de sistemas de bancos de dados NoSQL, realizar o seguinte procedimento sempre que recebermos uma requisição para o processamento e plotagem de um gráfico:

- Ao receber a requisição para gerar um gráfico de uma função específica (a escolha do grupo), para um conjunto de dados de entrada qualquer, nossa aplicação deve verificar se esse tipo de gráfico já foi plotado para o mesmo conjunto de dados de entrada.

- Caso não tenha sido: deverá ser gerada um índice que identifique unicamente o "conjunto de dados de entrada" + "tipo de gráfico requisitado", o gráfico deverá ser gerado pela função do R, e tanto o índice quanto o gráfico plotado deverão ser persistidos em um sistema de banco de dados NoSQL (a escolha do grupo).

- Caso tenha sido: ao invés de processar o conjunto de dados de entrada com o R para plotar o gráfico, iremos apenas recuperá-lo do local de armazenamento, nos livrando assim do overhead de processar os dados para plotar o gráfico novamente.

3 RESULTADOS

- Como resultado deste projeto de integração espera-se o seguinte:
  - Relatório técnico descrevendo o projeto do cache de plotagem de gráficos, contendo pelo menos as seguintes informações:
    - O projeto da solução (descrição global, algoritmo em pseudocódigo ou diagramação);
    - O sistema de banco de dados escolhido, com a justificativa da escolha;
    - Qual tipo de função (gráfico), escolhido para ser cacheado;
    - Qual a solução adotada para geração dos índices;
    - Em que formato os gráficos serão armazenados.
  - Adicionalmente devem ser implementados pelo menos 2 protótipos:
    - Um para o processamento dos gráficos em linguagem R;
    - Um para geração do índice e persistência dos gráficos (inserir e consultar).

O código dos protótipos deverá ser anexado ao relatório final.

4 CRONOGRAMA

Data	Atividade
13/02/2023 a 10/03/2023	Estudo das Unidades de Aprendizagem.
11/03/2023	Webconferência – Apresentação do Projeto Integrador.
12/03/2023 a 24/03/2023	Desenvolvimento do Projeto Integrador
25/03/2023	Webconferência – Ajustes do Projeto Integrador.
26/03/2023 a 04/04/2023	Desenvolvimento do Projeto Integrador
05/04/2023	Webconferência – Ajustes finais do Projeto Integrador.
06/04/2023 a 14/04/2023	Desenvolvimento do Projeto Integrador
15/04/2023	Apresentação do Projeto Integrador
15/04/2023	Entrega do Projeto Integrador

5 SUGESTÕES BIBLIOGRÁFICAS

Livros:

1. BOFF, ROCHA, M.; FERREIRA, P. G. Análise e exploração de sados com R. Lisboa: FCA. 2017.
2. WICKHAM, H.; GROLEMUND, G.; BATISTA, S. R para data science. Rio de Janeiro: Alta Books. 2019.
3. WICKHAM, H. Advanced R. 2. ed. New York: Chapman and Hall/CRC, 2019.
4. FOWLER, M.; SADALAGE, P. J. NoSQL essencial: um guia conciso para o mundo emergente da persistência poliglota. São Paulo: Novatec, 2013.
5. HOWS, D.; MEMBREY, P. Introdução ao MongoDB. São Paulo: Novatec, 2019.
6. MELO, A. B. de. Big Data e NoSQL: ontologias e estado da arte. Self Publishing. Amazon, 2020.

Complementar:

1. Kleppmann, Martin. Turning the database inside-out with Apache Samza. On-line: {https://www.confluent.io/blog/turning-the-database-inside-out-with-apache-samza/}.

Softwares:

- .