Programação Paralela e Distribuída

Odorico Machado Mendizabal

Universidade Federal de Santa Catarina – UFSC
Departamento de Informática e Estatística – INE

Projeto Final
Implementação usando Hadoop (Spark)

Objetivo

- Desenvolver um projeto completo de Big Data, onde os estudantes devem analisar os dados de uma base e extrair informações a partir destes dados

- Além da criatividade e metodologia de pesquisa utilizada, os estudantes devem explorar o framework de programação paralela Hadoop (ou Spark, se preferir) para proporcionar o processamento de dados distribuído

Trabalho

Análise de dados em larga escala

- Usando algum repositório com conjuntos de dados (data sets) – ex. Kaggle(https://www.kaggle.com/datasets), Dados.gov.br (http://dados.gov.br), etc.
    - Identificar um conjunto de dados de interesse
    - Elaborar os índices que serão avaliados
    - Descrever o seu estudo, com os resultados observados

Método

- Para os índices definidos em seu estudo, implemente um jobs de processamento do tipo MapReduce utilizando Hadoop ou Spark

Trabalho

Entrega

- Uma apresentação descrevendo:
  - O conjunto de dados estudado
  - Os índices pesquisados
  - Exemplos de implementação dos Jobs
  - Os resultados
- Entregar o código fonte

Trabalho – Apresentação

Formato

- Grupos podem ter no máximo 3 participantes
  - Todos devem participar na elaboração de conteúdo e na apresentação
  - Cada grupo terá 10 minutos para apresentação

Entrega

- De acordo com o cronograma da disciplina