Desenvolvendo uma aplicação para Apache Hadoop

EP 2 – Entrega 05/11/2012

Computação Paralela e Distribuída – MAC5742/MAC0431

Segundo Semestre de 2012

Prof. Alfredo Goldman (*gold at ime.usp.br*)

1. Introdução

Esse exercício procura explorar o potencial de paralelismo proposto pelo arcabouço Apache Hadoop. Hadoop é um arcabouço de código aberto, implementado em Java e utilizado para o processamento e armazenamento em larga escala, para alta demanda de dados, utilizando máquinas comuns. Os elementos chave do Hadoop são o modelo de programação MapReduce e o sistema de arquivos distribuído HDFS. O paradigma de programação MapReduce implementado pelo Hadoop se inspira em duas funções simples (Map e Reduce) presentes em diversas linguagens de programação funcionais. A função Map recebe uma lista como entrada, e aplicando uma função dada, gera uma nova lista como saída. Um exemplo simples é aplicar um fator multiplicador a uma lista, por exemplo, dobrando o valor de cada elemento:

map({1,2,3,4}, (x2)) > {2,4,6,8}

A função Reduce, similarmente à função Map, vai receber como entrada uma lista e, em geral, aplicará uma função para que a entrada seja reduzida a um único valor na saída. Algumas funções do tipo Reduce mais comuns seriam “mínimo”, “máximo” e “média”. Aplicando essas funções ao exemplo temos as seguintes saídas:

reduce({2,4,6,8}, mínimo) > 2

reduce({2,4,6,8}, máximo) > 8

reduce({2,4,6,8}, média) > 5

Esse EP será constituído de duas fases. Em uma primeira fase, deverá ser apresentado o problema que será implementado no arcabouço Hadoop. O desafio é mostrar o porque o problema se adapta à plataforma. Na segunda fase deverá ser entregue a implementação do problema para execução no arcabouço Apache Hadoop. O trabalho **pode** e **deve** ser feito por grupos de uma a duas pessoas. A entrega deve ser feita pelo Paca, através de um arquivo **ZIP** ou **TAR.GZ** com o **nome dos componentes**. Exemplos:

- `MarianaAlfredo.zip`
- `MarianaBravoAlfredoGoldman.tar.gz`

Nada de iniciais, por favor! Esse arquivo deve conter o **programa desenvolvido** (se for um projeto exportado do Eclipse, ajuda) e mais um **relatório** explicando detalhes sobre utilização e implementação do programa. **Note que se algum EPs estiver sem nome será descontada nota se o padrão não for seguido.**

As dúvidas devem ser resolvidas através do fórum da disciplina no Paca.

2. Fase 1 (Entrega 19/10/2012)

Nesta primeira etapa, deverá ser elaborada uma apresentação (duração máxima de 5 minutos), que além da caracterização do problema deverá ser conter também o esboço das funçoes Map e Reduce, ou seja, o que essas funções farão em sua aplicação e sobre qual conjunto de dados vão atuar. Deverá ser entregue um relatório descrevendo o problema e a abordagem para a resolução do mesmo, ilustrando as funçoes de Map e Reduce que serão implementadas no Apache Hadoop. Para as funções Map e Reduce deverão constar pelo menos um pseudo-código da abordagem proposta.

3. Fase 2 (Entrega 05/11/2012)

A aplicação deverá ser implementada de forma a funcionar no arcabouço Apache Hadoop. Para essa fase também deverá ser entregue um relatório descrevendo os detalhes de sua implementação em Java, especialmente das classes Map e Reduce.

4. Critérios de correção

Deverão ser entregues, além do relatório da fase 1 e dos códigos relativos à fase 2, um relatório descrevendo os detalhes da implementação. O relatório também deve conter os pontos positivos (facilidades) e negativos (dificuldades) encontrados durante o desenvolvimento da aplicação.