FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Sistemas de Computação em Cloud
Trabalho Prático 2

Devem manter os grupos do Trabalho 1 **Entrega: 7/Dez**

1. Introdução e Objectivos

O objectivo deste trabalho é duplo: a) Uma introdução ao ambiente Hadoop na *cloud* Azure, onde faz parte de um leque de produtos oferecidos sob o nome HDInsight; b) Uma introdução ao paradigma de programação Map-Reduce, aqui oferecido na implementação da Apache, designada Hadoop. Cada grupo terá de desenvolver uma palicação MR que processa um certo número de ficheiros e avalia (i) a escalabilidade da solução e (ii) também da plataforma.
Assim, no final do trabalho, deverá conseguir ter uma ideia aproximada do desempenho da sua aplicação e da forma como este é afectado pelo volume de dados (número e dimensão dos ficheiros processados) e pela dimensão do *cluster* Hadoop (i.e., como varia em função do número de nós dedicados à execução das tarefas, bem como em função dos recursos atribuídos a cada nó).

2. A aplicação

Considere os ficheiros do tipo WARC/WET (veja a aula prática correspondente), produzidos por um *crawler*, que guarda uma representação em texto "puro" das páginas Web visitadas. Para definir a relevância de um *site* podemos usar as mais variadas estatísticas, tais como os termos (palavras) mais comuns, o número e dimensão das páginas do *site*, ou o número de ligações (*links*) de outros *sites* para aquele que estamos a avaliar.

Versão-Base [Valorizada no máximo em 140 pontos.]: Crie uma aplicação MR que processa todos os ficheiros WARC/WET armazenados na directoria de *input* e produz uma "listagem" (i.e., cria um ficheiro de saída) na qual cada linha indica o site (S), o número de bytes (NB) extraídos pelo *crawler*, a duração da extração (T), e a "largura de banda" do site (obtida por NB/T). Assuma que cada *site* é representado por um único domínio DNS e que cada domínio só contém um *site*.

Opção [Valorizada em 60 pontos adicionais.]: Crie uma aplicação MR que processa os mesmos ficheiros WARC/WET e, para os 10 *sites* mais "volumosos" (ver nota abaixo), encontra as 10 palavras (com mais de 5 caracteres) mais frequentes, registando a sua frequência absoluta. **Nota:** processe apenas os *sites* com conteúdos no alfabeto Latino.

3. Sugestões

Siga as indicações das aulas práticas para criar um ambiente onde possa compilar os programas, antes de os submeter a um *cluster* HDInsight com um único *worker*; crie os nós com poucos recursos, para serem baratos. Note que criar o cluster demora uns 20 minutos, pelo que planeie o seu uso do tempo, e planeie os testes com antecedência…

Não se esqueça de destruir o cluster se acha que vai estar algumas horas sem o usar...

4. Documentação

Há muita coisa por aí; sugerimos que veja a API Hadoop da versão mais próxima da que vai usar no Azure (que é a 2.7.3) e os links fornecidos nas aulas práticas. É também aconselhável explorar as páginas do **`lemurproject`** em CMU e os ficheiros **`.java`** no zip fornecido nas aulas práticas.

5. Testes Obrigatórios e Entregáveis

O relatório deve incluir (a) uma descrição do problema, (b) o pseudo-código da solução, e (c) os aspectos mais relevantes da implementação (deve entregar um zip com o código fonte e incluir as instruções para a sua compilação e execução, bem como os links para os ficheiros que usou nos testes, de forma a que o seu trabalho possa ser "re-executado" …).

O relatório deve apresentar os resultados de desempenho, tanto em forma de texto/tabelas (condensados) como na forma gráfica (deve procurar a forma mais interessante/informativa para os exibir).

Tente avaliar o desempenho com, pelo menos, o seguinte conjunto de dados: a) um ficheiro; b) quatro ficheiros; c) oito ficheiros; d) 16 ficheiros. Cada conjunto deve ainda ser avaliado com um número apropriado de nós, começando por um nó *worker*; deve determinar qual o número final de nós, bem como o incremento que quer usar... Deve também fazer algumas experiências com o número de recursos por nó, tais como número de vCPUs, e RAM e tentar perceber, em cada caso, quantos *mappers* e *reducers* foram lançados.

Nota: planeie os testes de forma a que, quando cria um *cluster*, possa fazer o máximo de testes com essa configuração, senão o crédito pode esgotar-se rapidamente...