Aplicação MapReduce - Descrição

Esta aplicação MapReduce pretende receber um conjunto de dados na forma de log e processá-los de forma a ser possível responder eficientemente a 3 querys efectuadas ao webserver da aplicação Birdwatch:

- Dado uma data, mostrar a torre que observou o pássaro com a maior envergadura em períodos de chuva;
- Dado uma data e um identificador de torre mostrar o peso total de todos os passáros vistos por essa mesma torre;
- Listar todos os pássaros marcados que não foram vistos por mais de uma semana.

Mapper

Input -> tower-0, 2014-10-22, 00:02:07, eagle-74, 1222, 170, 3

Output:

No Mapper para cada entrada do log o output do Mapper será **dois pares key-value**. Para efeitos de processamento relativamente à **Query 1** e **Query 2** o output será o seguinte par key-value:

((date), (towerid, weight, birdid, wingspan))

Para efeitos de processamento relativamente à **Query 3** o output será o seguinte par key-value:

(birdid, date)

Para ser possível efectuar isto num só mapper criámos uma classe chamada **MapperOutputWritable** para englobar todos os valores possíveis para o value. Nesta classe temos uma flag **querytype**, onde se o seu valor for 0 então é um output correspondente a um processamento para a **query 1** e **query 2**, e se for 1 é um output correspondente a um processamento para a **query 3**.

Reducer

Input 1 -> (date, (towerid, weight, birdid, wingspan, querytype))

Input 2 -> (birdid, (date, querytype))

Output:

Tal como no **mapper**, o **reducer** pode devolver 3 possíveis formatos no **value** do output. No caso de no value do input o campo **querytype for igual a 0**, o **reducer** irá devolver os seguintes outputs:

Output 1 -> ((towerid, date), sum(weight))

Output 2 -> (date, max(wingspan))

Sendo que o **sum(weight)** equivale à soma do peso de todos os pássaros observados numa certa towerid numa certa data, e o **max(wingspan)** o pássaro com maior envergadura numa certa data. Estes valores são calculados somando todos os valores do campo weight para a mesma key (towerid, date) e verifincado qual o maior winspan para o a mesma key (date).

No caso de no value do input o campo **querytype for igual a 1**, o **reducer** irá devolver os seguintes outputs:

Output -> (birdid, lastseen)

O campo **lastseen** corresponde à última data onde o pássaro foi observado. Isto é calculado verificando para o mesmo birdid qual vai sendo a última data em que o pássaro foi observado, conforme se vão processando os values dessa key (birdid).