MapReduce

José A. Clemente

MapReduce

Framework que nos permite procesar de manera paralela una gran cantidad de datos.

Framework más usado para proceso de datos en Big Data. Soportado por:

- Hadoop
- Cassandra
- CouchDB
- MongoDB

MapReduce

Podemos implementarlo con JavaScript. No hace falta compilar nada. Los inicios de MapReduce fueron en Google (calcular page range de una página).
Se divide en 2 fases:

- Map. Aplicado en paralelo para cada uno de los registros a procesar. Agrupa por clave-valor.

- Reduce. Esta función pasando un parametro como clave, nos devuelve un listado de los valores por clave.

MapReduce

1. Map. Este proceso hace un split por cada una de las claves indicadas y devuelve un conjunto clave-valor.
2. Sort. El framework aplica de forma innata una ordenación.
3. Reduce. Procesa por cada una de las claves todos los valores pasados.

Nos permite procesar una gran cantidad de registros muy rapido.

MapReduce

```
{
    ‘cliente’: ‘c001’,
    ‘lineas’: [
        { ‘id’: ‘pantalon’, ‘unidades’: 4, ‘importe’: 10 },
        { ‘id’: ‘calcetin’, ‘unidades’: 2, ‘importe’: 6 }
    ]
}
```

Podemos ejecutar mapReduce o aplicarlo desde la shell o bien en un script JS (mapReduce.js,p. ej.).

MapReduce

```
var mapFunction = function() {
    for (var i = 0; i<= this.lineas.length; i++) {
        var key = this.lineas[i].id;
        var value = { subtotal: this.lineas[i].unidades };

        //Función emit para agregar un valor a la clave
        emit(key, value);
    }
}
```

Con el emit no se van guardando. Las va cogiendo al vuelo y ya las tiene. Me saca la clave-valor que me interesa (id y unidades). Podrian haber sido otras. Podemos aplicar MapReduce sobre nuestros conjuntos de datos tantas veces como queramos. Creando salidas nuevas (o colecciones cada vez).

MapReduce

```
var reduceFunction= function(id, countObjVals) {
        reduceVal = { total: 0 }

        for (var i = 0; i<= countObjVals.length; i++) {
                reduceVal.total += countObjVals[i].subtotal;
        }
        return reduceVal ;
}
```

Esta función incrementa el numero de unidades para el id especificado.

MapReduce

Falta ejecutarlo. Se puede hacer desde la shell o con el script creado. Por ejemplo, con esta función:

db.facturas.mapReduce(mapFunction, reduceFunction, { out: ‘map_reduce_result’ })

Se ejecuta sobre una colección (faccturas). Tercer parametro indica donde quiero el resultado. En este caso será una nueva colección llamada ‘map_reduce_result’

MapReduce

¿Cómo ejecuto el script?

mongo 127.0.0.1/facturas mapreduce.js

Si nos conectamos a la base de datos y hacemos un show collection, veremos la nueva colección.

Si quiero ver en detalle mi colección: db.map_reduce_result.find()