# Qual​ ​o​ ​objetivo​ ​do​ ​comando​ ​​cache​​ ​​​​em​ ​Spark?

O metodo cache() usa o nivel de armazenamento padrão, armazenando os objetos em memória (StorageLevel.MEMORY)

# O​ ​mesmo​ ​código​ ​implementado​ ​em​ ​Spark​ ​é​ ​normalmente​ ​mais​ ​rápido​ ​que​ ​a​ ​implementação​ ​equivalente​ ​em MapReduce.​ ​Por​ ​quê?

Porque Spark faz todo o seu processamento e armazenamento em memória, evitando uso desnecessário de disco.

# Qual​ ​é​ ​a​ ​função​ ​do​ ​​SparkContext​​?

É o principal ponto de entrada para as funcionalidades do spark. Ele representa a conectividade com o cluster spark e pode ser usado para criar RDDs, acumuladores e variáveis no cluster.

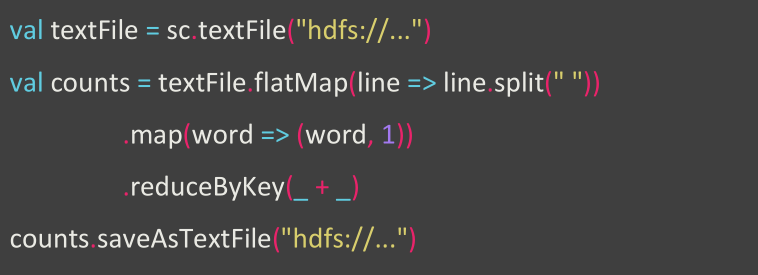
# Explique​ ​com​ ​suas​ ​palavras​ ​​ ​o​ ​que​ ​é​ ​​Resilient​​ ​​Distributed​​ ​​Datasets​​​ ​(RDD).

Representam coleções imutáveis e particionadas entre os nós de registros que podem ser processados em paralelo.

# GroupByKey​​ ​​​​é​ ​menos​ ​eficiente​ ​que​ ​​reduceByKey​​ ​​​​em​ ​grandes​ ​dataset.​ ​Por​ ​quê?

O groupbykey os dados são organizados de acordo com o valor de sua chave em diversos RDDs pelos nós da rede. Isso gera muito trafego em memória e em rede. O reducebykey agrupa e agrega o dado em cada partição dentro de cada nó antes de fazer qualquer tipo de transmissão.

# Explique​ ​o​ ​que​ ​o​ ​código​ ​Scala​ ​abaixo​ ​faz.



* Abre o arquivo de texto no hdfs;
* Divide as linhas pelos espaços;
* Atribui o valor 1 para cada palavra;
* Faz a agregação dos pares (palavra, 1)
* Salva o resultado (palavra – contagem ttl da palavra no arquivo) em um arquivo de texto no hdfs.