Contextualização

- O MapReduce surge como uma solução para computação paralela em grandes coleções de dados em clusters de computadores, contornando problemas como gestão de centenas de processadores diferentes, paralelização e a distribuição dessa elevada quantidade de dados, agendando e monitorizado estados de I/O, e a tolerância em casos de falhas.

(Cluster - Conjunto de computadores ligados à mesma rede, com especificações de hardware semelhantes)

- A imagem é o reflexo das grandes quantidades de informações com que as empresas são capazes de lida no dia-a-dia através do MapReduce, tendo sido a Google a primeira empresa a implementar este modelo de programação.

Funcionamento

MapReduce é um modelo de programação inspirado em 2 funções distintas, com o objetivo de processar e gerar uma grande quantidade de dados através de um algoritmo de paralelização e distribuição num cluster.

- Map

Responsável pela filtração e organização através de um par key/value de modo a gerar pares de keys e values intermédios.

- Reduce

Função responsável pelo agrupamento de todos os values intermédios com a mesma key intermedia, sendo posteriormente acedido pelos utilizador através de um iterador

permitindo a manipulação de listas demasiado grande para serem guardadas em memória.

- Shuffle

Responsável por uma redistribuição de informação com base nas keys de output resultantes da função Map, de modo a que toda a informação relativa a uma determinada key se encontre no mesmo nó de trabalho.

Vantagens do MapReduce

Na implementação de MapReduce em multi-thread verifica-se um performance substancialmente maior, assim como uma melhor escala e tolerância de falhas, quando comparado a implementações de MapReduce em single-thread e non-MapReduce.

Os programas em MapReduce são automaticamente paralelizados e executados num cluster de grandes dimensões. Deste modo o sistema de execução responsabiliza-se pelos detalhes da partição dos dados de input, o agendamento da execução do programa no conjuntos de maquinas, a manipulação de falhas das mesmas e a gestão da comunicação entre as mesmas.

Acima de tudo, permite que programadores possam fazer uso da utilização de recursos de um grande sistema de distribuição sem qualquer experiência anterior com sistemas paralelos ou distribuídos.

Monad

Um recurso de algumas linguagens como Clojure, Haskell e Scala, que permite que o MapReduce seja expresso de uma forma mais natural, através de técnicas de programação funcional.

A ideia básica consiste na conversão de cada estado map/reduce numa monadic function abstraindo peculiaridades semânticas da interface sintática.

Apache Hadoop

Apache Hadoop é um framework de software open-source para distribuição de armazenamento e distribuição de processamento em clusters de grande informação.

Faz a separação de ficheiros em blocos e distribui os mesmos entres os vários nos de um cluster.

Para o processamento da informação, o MapReduce da Hadoop gera ficheiros Jar e envia-os para os nos que iram processar essa informação em paralelo. Este modelo tem como vantagem a localização da data, em contraste com a arquitetura convencional baseada num sistema de ficheiro em paralelo.

Spark

Spark é um projecto da incubadora Apache, visto como sendo um projeto da próxima geração relativamente ao MapReduce, apostando no performance e na velocidade.

Spark proporciona uma API de alto nível para Scala, Java, Python com um conjunto de ferramentas de alto nível como SQL e processamento de estruturas de informação. Oferece também MLlib para machine learning, GraphX para o processamento de gráficos.