Nome: Joel Sepulveda Martins

O MapReduce é um modelo de programação que considera processamento  
distribuído, sendo amplamente utilizado pelo Apache Hadoop. Com base no seu  
conhecimento sobre MapReduce, responda as seguintes questões a seguir.  
1) O processamento com MapReduce envolve a execução de quatro etapas  
principais: splitting dos dados, map, sort/shuffle e reduce. Explique com suas  
palavras, as etapas de map e reduce.

R: O MapReduce é um modelo de programação que permite processar grandes quantidades de dados em paralelo e distribuído em um cluster de computadores. Ele é composto por quatro etapas principais: splitting dos dados, map, sort/shuffle e reduce. etapa de map, uma função de mapeamento distribui os dados em diversos nós de processamento e armazenamento. Essa função é responsável pelo processamento primário dos dados de entrada. Então, os resultados dessa fase são enviados para a função de redução como entradas. etapa de reduce, uma função agrega e sumariza os resultados obtidos no mapeamento, para gerar um resultado

2) Entre as quatro etapas do MapReduce citadas anteriormente, quais delas são  
de responsabilidade do desenvolvedor em termos de programação?

O desenvolvedor é responsável pelas etapas de map e reduce em termos de programação. Na etapa de map, o desenvolvedor deve escrever uma função que recebe um conjunto de dados e produz um conjunto de pares chave-valor intermediários. Na etapa de reduce, o desenvolvedor deve escrever uma função que recebe um conjunto de pares chave-valor intermediários e produz um conjunto menor de pares chave-valor

3) Quantas funções de map são executadas em paralelo em um único  
DataNode?

R: O número de funções de map que são executadas em paralelo em um único DataNode é determinado pelo número de slots de mapa disponíveis no DataNode. O número de slots de mapa é definido pelo administrador do cluster e pode ser configurado para cada nó individualmente

4) Explique os diferentes níveis de paralelismo possíveis durante o  
processamento com MapReduce.

R: Paralelismo de entrada: refere-se à capacidade de processar diferentes blocos de dados em paralelo.Paralelismo de map: refere-se à capacidade de processar diferentes registros dentro de um bloco de dados em paralelo.Paralelismo de reduce: refere-se à capacidade de processar diferentes grupos de registros em paralelo  
5) As etapas de map e reduce podem ser executadas ao mesmo tempo ou existe  
uma dependência entre as saídas geradas em uma das etapas para o início de  
outra etapa? Caso exista tal dependência, qual a vantagem do MapReduce em  
termos de paralelismo?

R: As etapas de map e reduce podem ser executadas ao mesmo tempo. Não há dependência entre as saídas geradas em uma das etapas para o início de outra etapa.A vantagem do MapReduce em termos de paralelismo é que ele permite que as tarefas sejam executadas em paralelo em diferentes nós do cluster, o que aumenta a eficiência e a velocidade do processamento.