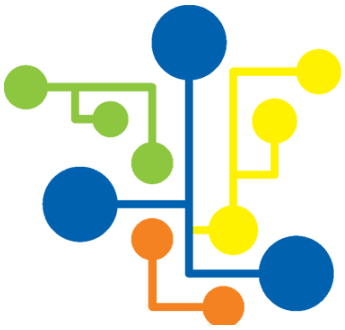

Data Science Academy

www.datascienceacademy.com.br

Data Lake – Design, Projeto e Integração

O Que é um Cluster?

Antes de falarmos sobre um cluster hadoop, vamos compreender o que é um cluster.

Quando o assunto é computação de alto desempenho, não é difícil pensarmos em servidores, sofisticados e caros respondendo por este trabalho. No entanto, é possível obter resultados tão bons quanto ou superiores a partir de alguma solução de **cluster** - uma tecnologia capaz de fazer computadores mais simples trabalharem em conjunto, como se formassem uma única máquina.

Mas o que é um cluster? Cluster (ou *clustering*) é, em poucas palavras, o nome dado a um sistema que relaciona dois ou mais computadores para que estes trabalhem de maneira conjunta no intuito de processar uma tarefa. Estas máquinas dividem entre si as atividades de processamento e executam este trabalho de maneira simultânea.

Cada computador que faz parte do cluster recebe o nome de *nó* (ou *node*). Teoricamente, não há limite máximo de nós, mas independentemente da

quantidade de máquinas que o compõe, o cluster deve ser "transparente", ou seja, ser visto pelo usuário ou por outro sistema que necessita deste processamento como um único computador.

Os nós do cluster devem ser interconectados, preferencialmente, por uma tecnologia de rede conhecida, para fins de manutenção e controle de custos, como a Ethernet. É extremamente importante que o padrão adotado permita a inclusão ou a retirada de nodes com o cluster em funcionamento, do contrário, o trabalho de remoção e substituição de um computador que apresenta problemas, por exemplo, faria a aplicação como um todo parar.

A computação em cluster se mostra muitas vezes como uma solução viável porque os nodes podem até mesmo ser compostos por computadores simples, como PCs de desempenho mediano. Juntos, eles configuram um sistema de processamento com capacidade suficiente para dar conta de determinadas aplicações que, se fossem atendidas por supercomputadores ou servidores sofisticados, exigiriam investimentos muito maiores. Não é necessário haver um conjunto de hardware exatamente igual em cada nó. Por outro lado, é importante que todas as máquinas utilizem o mesmo sistema operacional, de forma a garantir que o software que controla o cluster consiga gerenciar todos os computadores que o integram.

As tecnologias de Clustering possibilitam a solução de diversos problemas que envolvem grande volume de processamento. As aplicações que um cluster pode ter são diversas, indo desde a simples melhora no desempenho de um determinado sistema ou a hospedagem de um site, até o processo de pesquisas científicas complexas. O que realmente chama a atenção, é que todo o processamento pode ser feito de maneira que pareça ser um único computador dotado de alta capacidade. Assim, é possível que determinadas aplicações sejam implementadas em cluster, mas sem interferir no funcionamento de outras aplicações que estejam relacionadas.

A origem da denominação "cluster" não é clara, mas sabe-se que as primeiras soluções de processamento paralelo remontam à década de 1960, havendo, a partir daí, alguns princípios que hoje formam a base da ideia de clustering. O fato é que o passar do tempo não torna o conceito ultrapassado. Há um motivo especial para isso: os clusters se relacionam intimamente à otimização

de recursos, uma necessidade constante em praticamente qualquer cenário computacional. E este aspecto pode se tornar ainda mais atraente quando a ideia de cluster é associada a conceitos mais recentes, como Cloud Computing e Virtualização.