

O EMPREGO DE *CLUSTER* EM AMBIENTE DE COMPUTAÇÃO EM NUVEM PARA A ARMAZENAGEM E PROCESSAMENTO DE GRANDES QUANTIDADES DE DADOS (APOIO UNIP)

Aluno: Rodrigo Berto de Oliveira

Orientador: Prof. Dr. Jair Minoro Abe

Curso: Engenharia da Computação

Campus: Indianópolis

Na definição de Tenenbaum (2001), “Um sistema distribuído é uma série de computadores independentes que são visualizados pelos usuários como um sistema único”. Usuários, nessa definição, podem ser pessoas e sistemas que utilizam o *cluster* como parte do processamento ou armazenagem de dados. Cada computador do *cluster*, também chamado de nó ou nodo, pode ser um sistema multiprocessado ou não com memória, facilidades de I/O e um sistema operacional. Uma nuvem computacional provê um ambiente homogêneo (mesmo sistema operacional e mesmas bibliotecas em todos os nós da nuvem, proveniente da virtualização). Nuvens computacionais possibilitam um controle total sobre os recursos dedicados (em muitos casos, a nuvem é configurada de tal forma que os recursos computacionais necessários para as aplicações sejam quase que integralmente dedicados às aplicações que os utilizam; havendo necessidade de mais recursos, podem ser adicionados conforme as necessidades das aplicações). Com a combinação de nuvens de computadores, virtualização de servidores e *clusters* de computadores foram criados ambientes de computação que podem produzir sistemas de armazenamento e processamento de grandes quantidades de dados em computadores e servidores convencionais. Eles cumprem o objetivo por meio do emprego de técnicas tais como arranjos redundantes de discos independentes, replicação de arquivos e somas de verificações. O termo *Big Data* é a terminologia que vem sendo empregada para essa área da computação que une sistemas distribuídos (*clusters*), computação em nuvem, armazenamento escalável, análise e previsão de

dados. Em 2003, os pesquisadores Sanjay Guemawat, Howard Gobioff e Shun-Tak Leung, da empresa *Google*, publicaram um artigo sobre o modelo de armazenamento e processamento de grandes quantidades de dados, que acabou sendo a base para o desenvolvimento de sistemas de código aberto para a manipulação e armazenamento de grandes quantidades de dados. Um dos projetos mais maduros utilizado para esse fim é o *Hadoop*, mantido pela Fundação Apache.