This paper considers the joint fronthaul compression and transmit beamforming design for the uplink cloud radio access network (C-RAN), in which multi-antenna users communicate with a cloud-computing based centralized processor (CP) through multi-antenna base-stations (BSs) serving as relay nodes. A compress-and-forward relaying strategy, named the virtual multiple-access channel (VMAC) scheme, is employed, in which the BSs can either perform single-user compression or Wyner-Ziv coding to quantize the received signals and send the quantization bits to the CP via capacity-limited fronthaul links; the CP performs successive decoding with either successive interference cancellation (SIC) receiver or linear minimum-mean square-error (MMSE) receiver. Under this setup, this paper investigates the joint optimization of the transmit beamformers and the quantization noise covariance matrices for maximizing the network utility. A novel weighted minimum-mean-square-error successive convex approximation (WMMSE-SCA) algorithm is first proposed for maximizing the weighted sum rate under the user transmit power and fronthaul capacity constraints with single-user compression. Assuming a heuristic decompression order, the proposed algorithm is then adapted for optimizing the transmit beamforming and fronthaul compression under Wyner-Ziv coding. This paper also proposes a low-complexity separate design consisting of optimizing transmit beamformers for the Gaussian vector multiple-access channel along with per-antenna quantizers with uniform quantization noise levels across the antennas at each BS. Numerical results show that majority of the performance gain brought by C-RAN comes from the implementation of SIC at the CP. Furthermore, the low complexity separate design already performs very close to the optimized joint design in regime of practical interest.