Predictive resource allocation is an emerging approach to improve the performance of mobile systems as human behavior is reported predictable by leveraging big data analytics. Yet what information can be predicted by big data, what information need to be predicted for wireless access optimization, how to translate the information, and how to exploit the synthetic knowledge for allocating radio resources are not well understood and largely explored. In this paper, we are concerned with the latter two issues. In particular, we devise an energy-saving resource planning and allocation policy for multiple base stations (BSs) to serve mobile users with non-real-time (NRT) traffic by exploiting the user, network, and application levels of context information, where RT traffic may occupy partial resources of each BS. Inspired by the solution from an energy minimization problem with future instantaneous information, a low complexity multi-timescale predictive policy is proposed. Upon the arrival of each NRT user request, the resource planning is made with the user and network level context information, defined as the average channel gains of the NRT users and the statistics of residual bandwidth after serving RT traffic, with which the scheduling, power allocation, and BS sleeping can be accomplished after instantaneous channel information and residual network resource are available at each BS in each time slot. Simulation results show that the proposed policy can dramatically reduce the energy consumed by the BSs for serving the NRT traffic.