Diplo Cloud: Efficient and Scalable Management of RDF Data in the Cloud

Despite recent advances in distributed RDF data management, processing large-amounts of RDF data in the cloud is still very challenging. In spite of its seemingly simple data model, RDF actually encodes rich and complex graphs mixing both instance and schema-level data. Sharding such data using classical techniques or partitioning the graph using traditional min-cut algorithms leads to very inefficient distributed operations and to a high number of joins. In this paper, we describe Diplo Cloud, an efficient and scalable distributed RDF data management system for the cloud. Contrary to previous approaches, Diplo Cloud runs a physiological analysis of both instance and schema information prior to partitioning the data. In this paper, we describe the architecture of Diplo Cloud, its main data structures, as well as the new algorithms we use to partition and distribute data. We also present an extensive evaluation of Diplo Cloud showing that our system is often two orders of magnitude faster than state-of-the-art systems on standard workloads.