Survey of machine learning algorithms on Spark over DHT-based Structures

Spyros Sioutas, Phivos Mylonas, Alexandros Panaretos,
Panagiotis Gerolymatos, Dimitrios Vogiatzis, Eleftherios Karavaras, and
Thomas Spitieris

Department of Informatics, Ionian University,
49100 Corfu, Greece
{sioutas, fmylonas, alex, c12gero, p12vogi, p12kara, p12spit}@ionio.gr

Abstract. Over the past few years there have been proposed many solutions on data storage, data management and data retrieval systems. These solutions can process massive amount of data stored in relational or distributed database management systems. In addition, decision making analytics and predictive computational statistics are some of the most common and well studied fields in computer science. In this paper, we demonstrate the implementation of machine learning algorithms over an open-source distributed database management system that can run in parallel on a cluster. In order to accomplish that we propose a system architecture scheme such as Apache Spark over Apache Cassandra. This paper presents a survey of the most common machine learning algorithms and the results of the experiments performed over a point of sales data set.

Keywords: machine learning· Spark· Cassandra· DHT-based structures