CoPart: A Context-based Partitioning Technique for Big Data

Sara Migliorini (sara.migliorini@univr.it)
Universita degli Studi di Verona https://orcid.org/0000-0003-3675-7243
Alberto Belussi
University of Verona: Universita degli Studi di Verona
Elisa Quintarelli
University of Verona: Universita degli Studi di Verona
Damiano Carra
University of Verona: Universita degli Studi di Verona

Research

Keywords: Big data, partitioning technique, contextbased queries

Posted Date: October 14th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-91158/v1

Version of Record: A version of this preprint was published on January 19th, 2021. See the published version at https://doi.org/10.1186/s40537-021-00410-4.

Abstract

The MapReduce programming paradigm is frequently used in order to process and analyse huge amount of data. This paradigm relies on the ability to apply the same operation in parallel on independent chunks of data. The consequence is that the overall performances greatly depend on the way data are partitioned among the various computation nodes. The default partitioning technique provided by systems like Hadoop or Spark, basically performs a random subdivision of the input records, without considering the nature and correlation between them. Even if such approach can be appropriate in the simplest case where all the input records have to be always analysed, it becomes a limit for sophisticated analyses that imply correlations between records that can be exploited to preliminary prune unnecessary computations.

In this paper we propose a partitioning technique which exploits the notion of context for partitioning data. We design a context-based multi-dimensional partitioning technique, called \copart, which considers not only the correlation of data w.r.t. contextual attributes, but also the distribution of each contextual dimension in the dataset. We experimentally compare our approach with existing ones, considering both quality criteria and the query execution times.

Full Text

This preprint is available for download as a PDF.

Figures

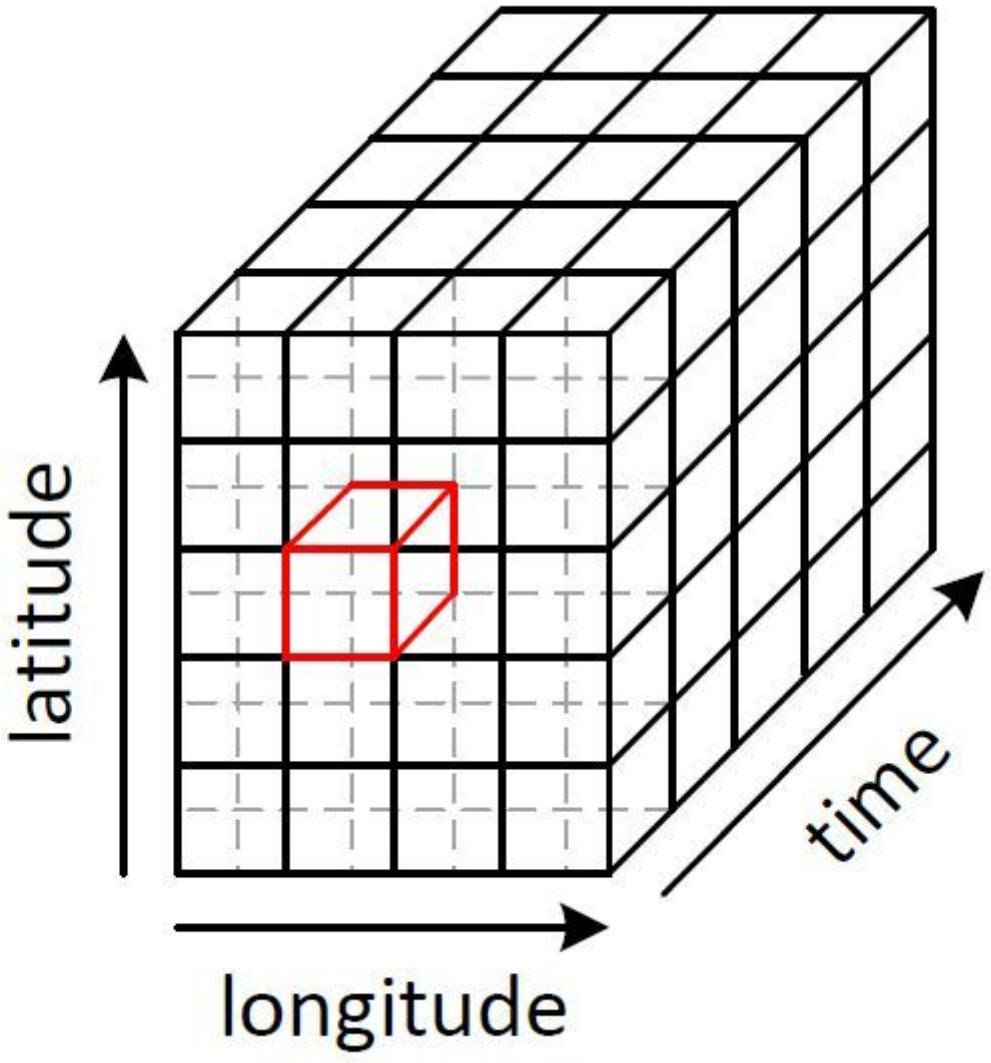

Figure 1

Example of multi-dimensional partitioning.

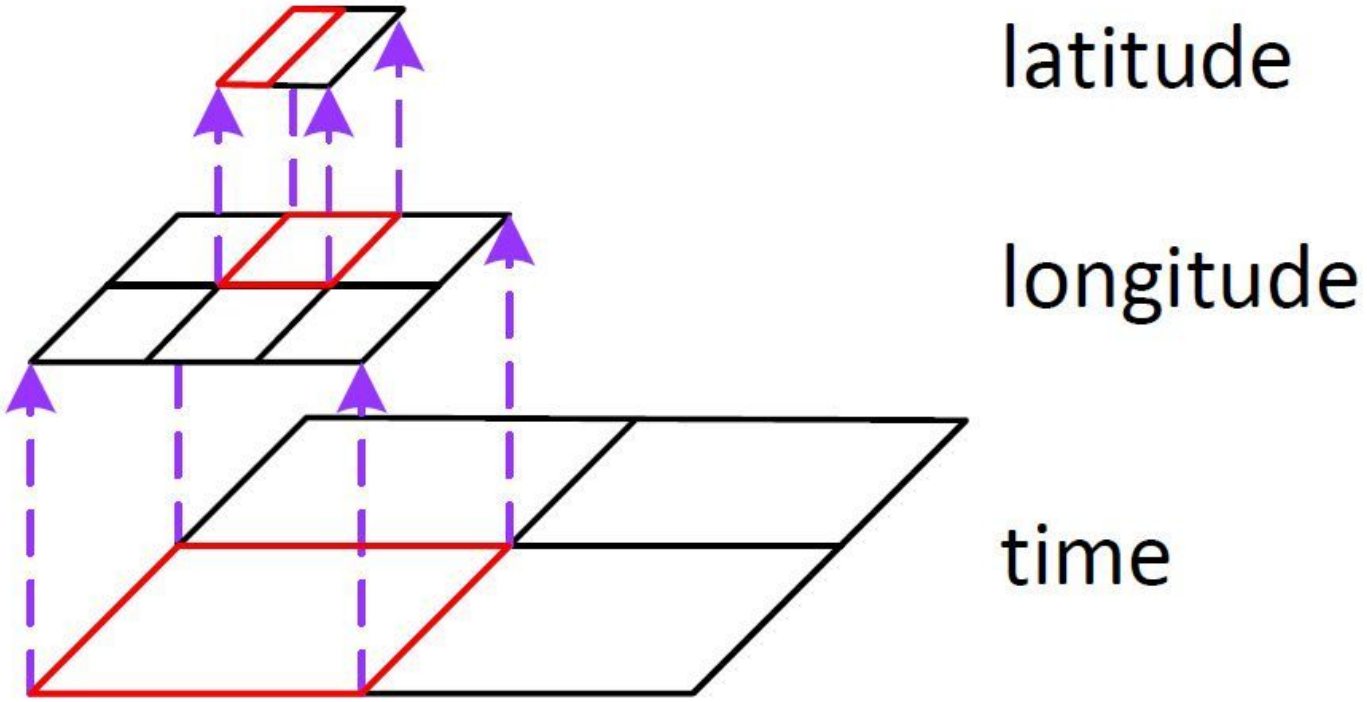

Figure 2

Example of multi-level partitioning.

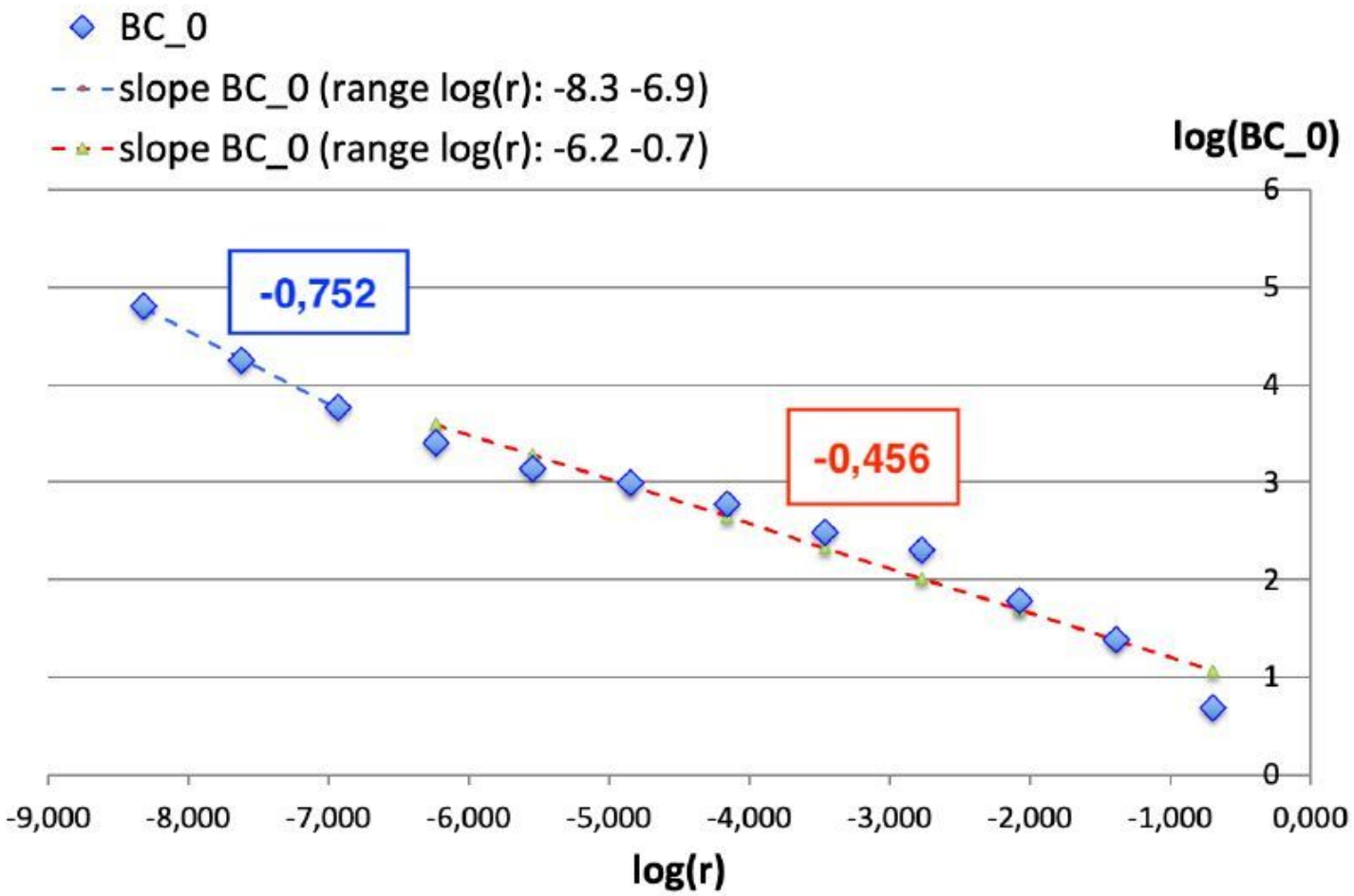

Figure 3

Example of box-counting plots (BC0r (D; long)) for the context attribute representing the longitude with q = 0. Notice that, in order to extract the behavior of the dataset on the largest scale range, the computation of the slope takes into account the presence of variations in the sequence of values. At the end, the slope of the straight line with maximum support (the red one in this case) is chosen.

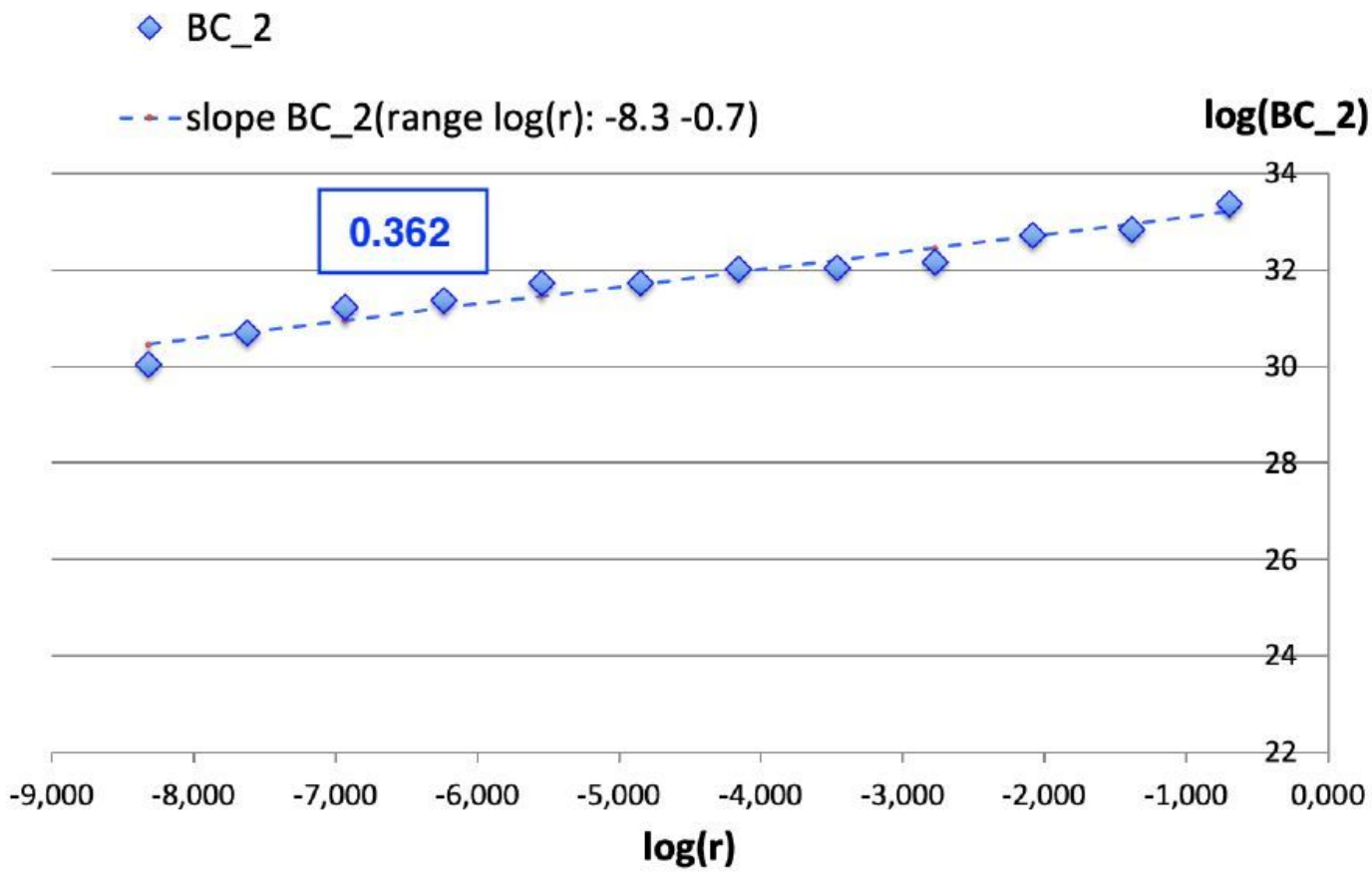

Figure 4

Example of box-counting plots (BC2 r (D; long)) for the context attribute representing the longitude with q = 2.

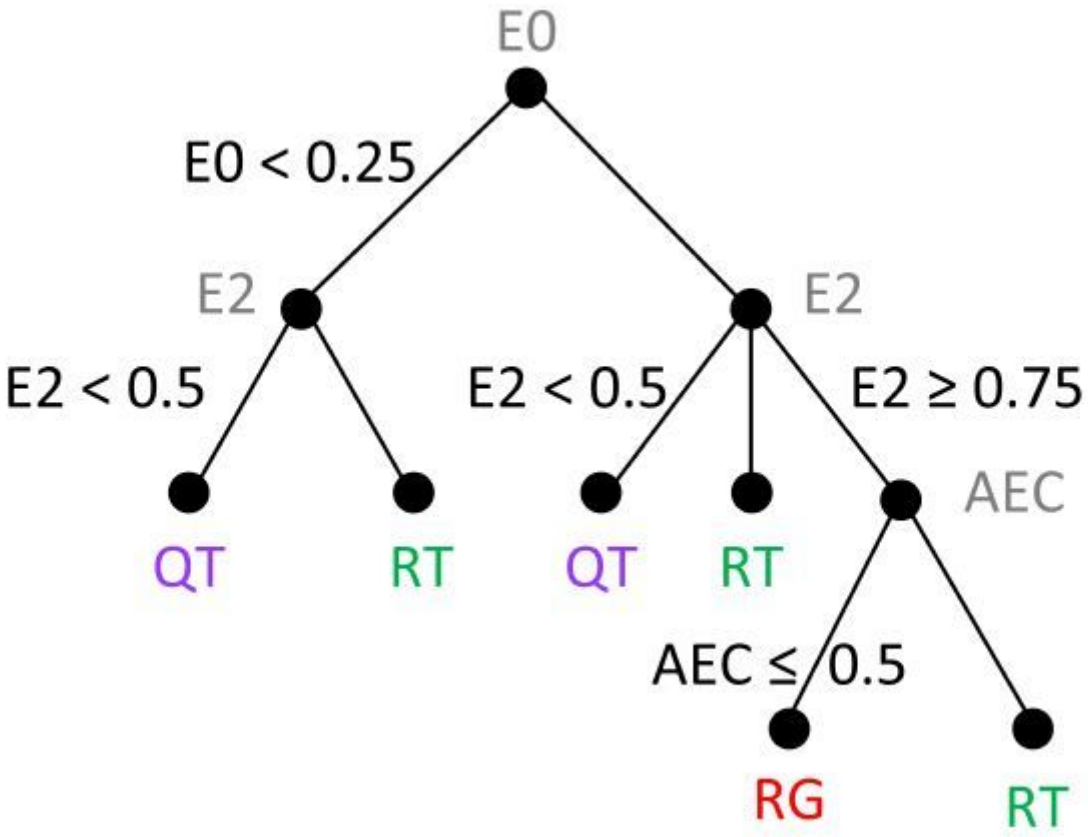

Figure 5

Decision tree