

# CoPart: A Context-based Partitioning Technique for Big Data

Sara Migliorini (✉ [sara.migliorini@univr.it](mailto:sara.migliorini@univr.it))

Università degli Studi di Verona <https://orcid.org/0000-0003-3675-7243>

Alberto Belussi

University of Verona: Università degli Studi di Verona

Elisa Quintarelli

University of Verona: Università degli Studi di Verona

Damiano Carra

University of Verona: Università degli Studi di Verona

---

## Research

**Keywords:** Big data, partitioning technique, contextbased queries

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

**License:** © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

---

## 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

Due to technical limitations, full-text HTML conversion of this manuscript could not be completed. However, the manuscript can be downloaded and accessed as a PDF.