SPARKSTAC: SPATIO-TEMPORAL CLIMATE DATA ANALYTIC FRAMEWORK

Madusha Jayawardhana

A dissertation submitted in partial fulfilment of the requirement for Master of Science degree in Advanced Software Engineering

Department of Computing
Informatics Institute of Technology, Sri Lanka
in collaboration with
University of Westminster, UK

2020

Abstract

Climate changes experienced worldwide is one of the major challenges faced by the global community nowadays. Large volumes of climate data from different sources reveal information about various climate aspects and analyzing those data would help in revealing hidden patterns within data to identify the correlations between various climate parameters and discover knowledge and insights. The exponential growth of climate data is making it a big data domain and creating new opportunities in climate science. However, the unprecedented growth of climate data is posing challenges to efficiently manage and analyze big climate data to conduct analytical operations and gain insights on climate changes. Climate data are of spatio-temporal in nature, consisting components in both spatial and time dimensions. The complex nature of the climate data and analytic algorithms make it difficult to implement an efficient way of analyzing climate data. Achieving that will require effective data management strategies, data parallelization and parallel execution of complex computing algorithms.

Apache Spark is widely accepted in big data domain because of its fast, in-memory distributed data processing ability which is much faster than Hadoop. However, Spark or Hadoop does not provide any native support for spatio-temporal or spatial data where users need to implement spatio-temporal operations incompetently by themselves. Furthermore, Spark cannot use data parallelization with optimal partitioning for spatio temporal distribution of data. A spatio temporal data analytic framework that can perform on demand climate data analytic operations efficiently is a timely need required by the climate scientists.

The developed system SparkSTAC is using climate datasets with variations of climate parameters with space and time components as inputs, and provides an in-memory , distributed big data analytic framework that can manage and analyze climate data more effectively and efficiently by using the data parallelization techniques that exploit spatial and temporal proximity of the data and partition pruning in query execution. With SparkSTAC, spatio temporal support is added to Spark together with spatio temporal partitioning strategy, indexing and operators like join, filter and KNN. Furthermore, SparkSTAC smoothly integrates with Spark without giving much

difference to Spark data analytic user. Experimental evaluation results show that SparkSTAC performs different spatio temporal analytic operations with high performance by using spatio temporal data locality.

Keywords: Big Data, Climate data analysis, Spatio-temporal, Spark, Partitioning