Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 119 (2017) 334–340

Procedia Computer Science

www.elsevier.com/locate/procedia

6th International Young Scientists Conference in HPC and Simulation, YSC 2017, 1-3 November 2017, Kotka, Finland

High-performance meteorological data processing framework for real-time analysis and visualization

Gali-Ketema Mbogo[a,*], Stepan V. Rakitin[a], Alexander Visheratin[a]

[a] *ITMO University, Saint-Petersburg, Russia*

Abstract

The current level of geoscience advancements enables analysis of the Earth as a highly complex system with many components, like hydrosphere, lithosphere, and atmosphere, which interact within and with others. Large amounts of observational data and simulation data generated by numerical models need to be analyzed in order to extract valuable knowledge from them. Visualization is a very important part of data analysis since it provides an easy and fast way to assess the data and understand, which parts of it are of particular interest and whether there are any errors in the dataset. The most efficient type of tools for meteorological data visualization are geographic information systems (GIS) because they provide great functionality for manipulating geospatial data. In this paper, we describe the processing framework, which can be efficiently used as a backend for GIS by providing rapid access to the data located at the remote storage nodes. Experimental results demonstrate that developed framework allows real-time data access and can successfully handle a large number of simultaneous requests.

Peer-review under responsibility of the scientific committee of the 6th International Young Scientist conference in HPC and Simulation

Keywords: GIS, meteorological data, NetCDF, high-performance computing

1. Introduction

Geoscience plays a crucial role in our life because a large number of areas of life nowadays strongly depend on it - maritime industry, mining industry, agriculture, city management, etc. Progressing advances in science and technology make it possible to perform complex analysis and very precise forecasts of meteorological and hydrological processes. The prime example of storage and processing enormous amounts of geoscience data is NASA, whose Center for Climate Simulation contains more than 32 petabytes of data. This data can be used as input for numerical models, which perform detailed simulations of weather or ocean conditions for deep analysis of the past events or for forecasting of meteorological phenomena in the nearest future.

[*] Corresponding author
E-mail address: ketema.galy@gmail.com

1877-0509 © 2018 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the scientific committee of the 6th International Young Scientist conference in HPC and Simulation
10.1016/j.procs.2017.11.192

There are three main formats for storing meteorological data [8] - NetCDF [14], GRIB[1], and HDF5 [7]. Although every format has its strengths and weaknesses, for our research we have chosen NetCDF data format. The first reason for this is the fact that NetCDF is very popular among scientific and government organizations - today more than 1300 organizations use it to store their data[2]. The second reason is peculiarities of the format. NetCDF is array-oriented binary format by design. The user can declare variables, dimensions, and attributes in a file. Variables usually hold continuous data and can be referred by their dimensions. Internally all arrays in NetCDF files are flat, which means that even if the variable data has multiple dimensions array containing it is still linear. By that knowing the dimensions of the variable, we can extract value from it using binary offsets without reading the whole array. Variables themselves can also be extracted from the file by offsets, which are specified in the header of the file. Described properties greatly increase the number of use cases of the format because they allow working with variables and values without reading the entire file into memory, thus making possible to create large files storing a lot of interconnected data.

Since NetCDF format is very general and gives the user a plenty of options how to organize the file structure, the conventions for climate and forecast (CF) metadata Eaton et al. [6] were developed to unify the way of processing meteorological data. According to these conventions, there are four types of coordinates: latitude, longitude, vertical level, and time. Schematic representation of data in CF format is presented in Figure 1. Variables may not have all variables, e.g., observational data from weather stations, which contain only time dimension.

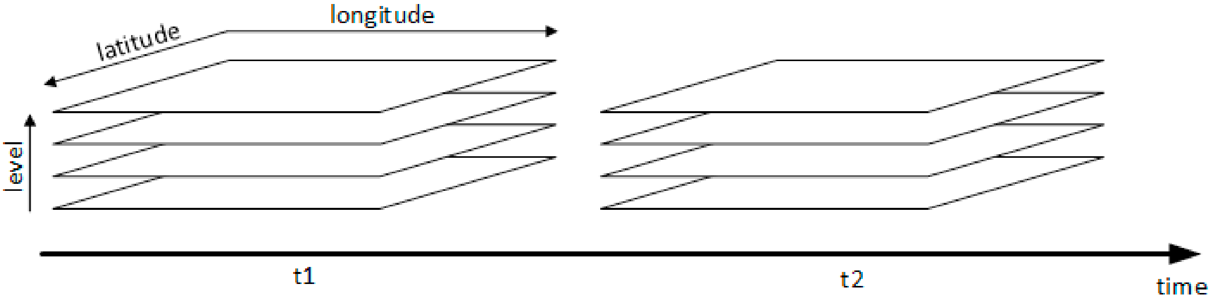

Fig. 1. Schema of variables representation in NetCDF format

When working with meteorological data, the very natural first step is data visualization, which helps to understand the data better. Today a large number of geographic information systems exist. These systems are able to display the data in an easy-to-use way and perform basic analytical operations, like mean calculation or outliers detection. However, even the most developed and widespread platforms, like ArcGIS or QGIS, are aimed at working with separate files and cannot provide processing of a large amount of data from different sources. The core problem, which prevents modern GIS from this functionality, is the lack of backend that could be capable of meeting data representation requirements of GIS platforms. The general problem of efficient storage and processing of meteorological data motivates scientists to develop novel approaches for this task. Authors of [16] developed a framework, which makes it possible to create scalable data analysis applications that can work with the popular scientific data formats SciMATE. This framework is based on the MATE system [9]. SciMATE provides MapReduce like API that allows users to create complex scientific applications and extend the framework for any data format. Experiments demonstrate that the developed system is functional and scales well on a multi-core cluster. Nevertheless, due to the complex nature of the SciMATE, it cannot be easily adapted to provide data in a real-time mode.

Duque-Mndez et al. in [5] proposed a model for a data warehouse in a star schema that allows online analytical processing of the data. Authors use two networks of hydro-meteorological stations as data providers. The paper presents a novel model of data storage and a number of use cases when this model can show better results than other approaches. The main disadvantage of the article is that authors did not present a developed system with experimental evaluation. Authors of [11] propose a scientific workflow framework for big geoscience data analytics. The proposed

[1] http://www.wmo.int/pages/prog/www/DPS/FM92-GRIB2-11-2003.pdf
[2] http://www.unidata.ucar.edu/software/netcdf/usage.html