УДК 004.67+ 004.62

И.В. Морозов, Р.А. Родригес Залепинос
Донецкий национальный технический университет, г. Донецк
кафедра компьютерных систем мониторинга

ОБЗОР ИНСТРУМЕНТОВ ОБРАБОТКИ ДАННЫХ ЭКОЛОГИЧЕСКОГО МОНИТОРИНГА

Аннотация

Морозов И.В., Родригес Залепинос Р.А. Обзор инструментов обработки данных экологического мониторинга. *Выполнен обзор инструмента обработки данных экологического мониторинга на примере NCO (NetCDF Common Operators). Описаны основные возможности инструмента NCO. Показаны примеры обработки данных экологического мониторинга.*

Ключевые слова: *методы обработки данных, интерфейсы, формат, файл, операторы.*

Постановка проблемы. Вопросы экологии в наше время принимают первостепенное значение. В мировой практике накоплен колоссальный опыт в области отслеживания и анализа вредных антропогенных воздействий на среду обитания. Современные технологии позволяют определять проблемы экологии еще на стадии их возникновения. Двумя основными методами отслеживания воздействия на нашу среду обитания и происходящих в ней изменениях являются мониторинг и контроль. Для сбора данных существует множество технических решений, однако, для обработки этих данных довольно мало инструментов. Поэтому в данной статье будет выполнен обзор бесплатного инструмента для обработки данных экологического мониторинга NCO (NetCDF Common Operators).

Цель статьи – провести обзор инструмента для обработки данных экологического мониторинга NCO, описать его основные возможности и показать примеры обработки данных.

Введение. NCO представляет собой набор программ, известных как операторы. Каждый оператор является автономным, программа командной строки выполняется на командном уровне. Операторы принимают NetCDF файлы на вход (в том числе HDF5 файлы, построенные с использованием API NetCDF), выполняют операцию (например, нахождение среднего) и создают NetCDF файл в качестве выходного файла. Формат NetCDF является машинно-независимым, самоописываемым, и открытым, облегчая тем самым множество проблем, связанных с доступом.

Набор утилит командной строки, называемый NetCDF operators (NCOs) доступен на большинстве машин, работающих под Linux, Mac и PC, как 32-

так и 64-разрядных платформах. Поддержка предоставляется для компиляции нативных исполняемых файлов Windows, используя Microsoft Visual Studio 2010 Compiler, или Cygwin. Как и все исполняемые файлы, NetCDF operators могут быть построены с использованием динамического связывания. Это уменьшает размер исполняемого файла и может значительно повысить производительность на многопользовательских системах.

NCO позволяет выполнять простые расчеты и действия с NetCDF или HDF4 файлами с минимальными знаниями NetCDF файлов. NCO может обрабатывать данные на порядок быстрее, чем Matlab или другие пакеты для анализа данных. В Национальном центре атмосферных исследований был разработан NCAR Command Language - подобный набор утилит со схожей функциональностью [1].

Операторы, облегчающие анализ данных, содержатся в самоописываемом формате NetCDF, доступны по ссылке http://unidata.ucar.edu/packages/NetCDF. Хотя большинство пользователей участвуют в научных исследованиях, эти форматы данных носят общий характер и одинаково полезны в различных областях, от сельского хозяйства до зоологии [2].

NetCDF (Network Common Data Form) представляет собой набор интерфейсов для доступа к массиво-ориентированным данным и свободно распространяемой коллекции библиотек для C, FORTRAN, C++, Java, и других языков. Библиотеки NetCDF поддерживают машинно-независимый формат для представления научных данных. Вместе, интерфейсы, библиотеки и формат поддерживают создание, доступ и совместное использование научных данных.

NetCDF данные [3]:

- *Самоописывающиеся*. Файл NetCDF включает в себя информацию о содержащихся в нем данных.
- *Портативные*. К NetCDF файлу могут обращаться компьютеры с различными способами хранения целых чисел, символов и чисел с плавающей точкой.
- *Масштабируемые*. К небольшой части большого набора данных может быть эффективно получен доступ.
- *Присоединяемые.* Данные могут быть присоединены к правильно структурированному NetCDF файлу без копирования набора данных или переопределения его структуры.
- *Параллельно доступные*. Один записывающий клиент и несколько читающих клиентов могут одновременно получить доступ к одному и тому же файлу NetCDF.
- *Обратносовместимые*. Доступ ко всем ранним формам NetCDF данным будет поддерживаться текущими и будущими версиями программного обеспечения.

Программное обеспечение NetCDF разработано в Боулдере, штат Колорадо, при участии многих пользователей NetCDF. Неполный список

организаций, использующих NetCDF для архивирования и доступа к своим данным: NOAA, EUMETSAT Data Centre, GOES-R Ground Segment Products, DOE/PCMDI CMIP5 (результаты моделирования климата), NASA/JPL, NASA/GSFC, National Center for Atmospheric Research (NCAR), Австралийский центр метеорологических и климатических исследований (CAWCR).

Формат файлов. Существует 4 формата NetCDF: классический формат, 64-бит смещенный формат, NetCDF-4 формат, NetCDF-4 классический формат модели. Структура классического и NetCDF-4 форматов представлена на рисунках 1 и 2 [4].

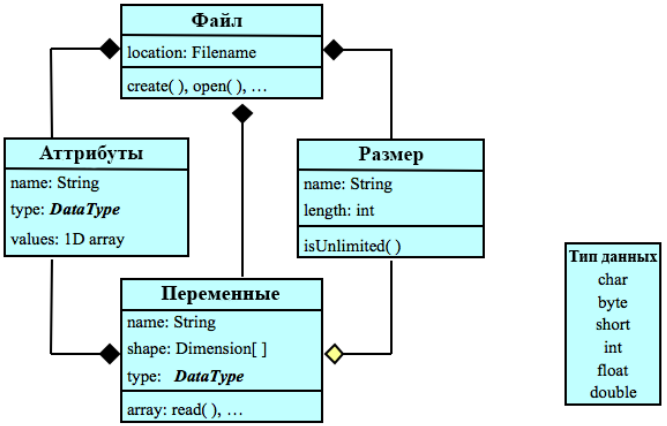

Рисунок 1 – Структура классического формата

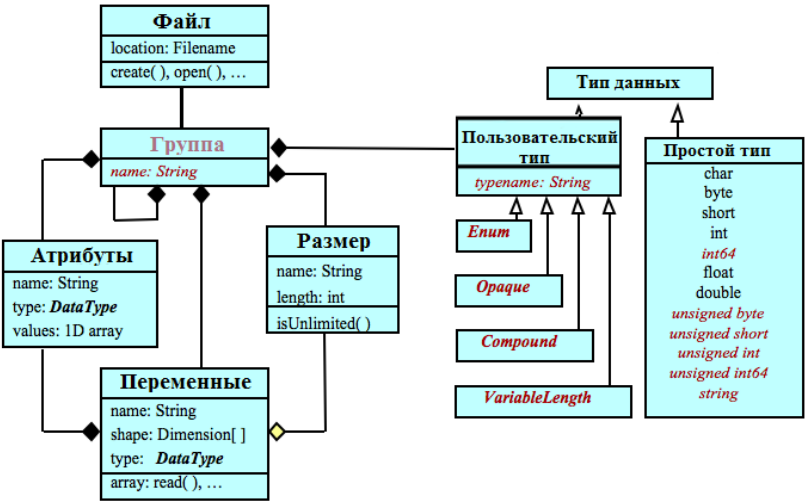

Рисунок 2 – Структура NetCDF-4 (Enhanced Data Model) формата

Поддержка файлов больших размеров (LFS) относится к операционным системам и С-библиотекам объектов для поддержки файлов, размером более 2 Гб. В некоторых 32-битных платформах размер смещения файла по умолчанию остается 4-байтовым целым числом, что ограничивает максимальный размер файлов до 2 Гб. Использование интерфейсов LFS и 64-битного типа смещения, максимальный размер может быть 2^{63} байт или 8 эксбибайт. Для некоторых современных платформ должны быть установлены макросы для поддержки файлов больших размеров или соответствующие флаги компилятора, чтобы построить библиотеку с поддержкой файлов больших размеров.

NetCDF распространяется с интерфейсами для C, Fortran77, Fortran90, и C++. Другие языки для которых существуют интерфейсы: Ada, IDL, Java, MATLAB, Perl, Python, R, Ruby, Tcl/Tk.

Примеры. Простые примеры Bash shell сценариев показывают, как усреднить данные в разных структурах. Здесь включены месячные, сезонные и среднегодовые средние с ежедневными или месячными данными в одном или нескольких файлах. Рассмотрим некоторые из них [4].

Вычисление суточных изменений по годам в одном файле. Предположим, что у нас есть ежедневные данные с 1 января 1990 по 31 декабря 2005 года в файле in.nc где поле времени имеет название time. На рисунке 3 показаны суточные изменения по годам в одном файле.

```
Месячное среднее:
for yyyy in {1990..2005}; do      # Цикл лет
  for moy in {1..12}; do          # Цикл месяцев
    mm=$( printf "%02d" ${moy} )  # Переход в 2-х значный формат
    # Среднее месяца yyyy-mm
    ncra -O -d time,"${yyyy}-${mm}-01","${yyyy}-${mm}-31" \
         in.nc in_${yyyy}${mm}.nc
  done
done
# Объединение ежемесячных файлов вместе
ncrcat -O in_??????.nc out.nc
Годовое среднее:
for yyyy in {1990..2005}; do      # Цикл лет
  ncra -O -d time,"${yyyy}-01-01","${yyyy}-12-31" in.nc in_${yyyy}.nc
done
# Объединение ежегодных файлов вместе
ncrcat -O in_????.nc out.nc
```

Рисунок 3 – Суточные изменения по годам в одном файле

Опция −O нужна для перезаписи уже существующих файлов. NCO поддерживает UDUnits, чтобы у пользователя была возможность использовать читаемые даты как измерение времени.

Суточные изменения по годам в одном файле. Внутри входного файла in.nc, записано время с января 1990 по декабрь 2005 года. На рисунке 4 показано вычисление сезонного и годового среднего.

```
Сезонное среднее:
ncra -O --mro -d time,"1990-12-01",,12,3 in.nc out.nc
Годовое среднее:
ncra -O --mro -d time,,,12,12 in.nc out.nc
```

Рисунок 4 – Ежемесячные данные в одном файле

Здесь используется функция подциклом (т.е. число после 4-ой запятой: «3» в сезонном примере и второе «12» в ежегодном), чтобы получить группу записей, разделенных регулярными интервалами. Опция –mro является переключателем ncra для Multi-Record Output вместо Single-Record Output.

Расчет толщины.

```
ncea -F -d level,8,8 hgt.mon.mean.nc hgt300.mon.mean.nc
ncea -F -d level,3,3 hgt.mon.mean.nc hgt850.mon.mean.nc
```

Рисунок 5 – Чтение данных для разных уровней давления

Повторный анализ NCEP / NCAR поставляется с данными для 17 вертикальных уровней. Берутся данные для уровней 300 и 850 миллибар давления. -F означает использовать FORTRAN-индексацию (нумерация начинается с 1). C-индексация начинается с 0. На рисунке 6 происходит вычитание hgt850 и hgt300, и запись толщины.

```
ncdiff hgt300.mon.mean.nc hgt850.mon.mean.nc
thickness300850.mon.mean.nc
```

Рисунок 6 – Расчет толщины

Выводы. Проведен обзор инструмента обработки данных экологического мониторинга NetCDF Common Operators. Описаны основные возможности программы. Показаны некоторые примеры работы инструмента NetCDF. Обзор показал, что NCO довольно мощный и гибкий инструмент для анализа и обработки данных экологического мониторинга.

Список литературы

1. Wang, D. L., C. S. Zender, and S. F. Jenks, Efficient Clustered Server-side Data Analysis Workflows using SWAMP, Earth Sci. Inform., Vol. 2(3), P. 141 – 155, 2009.
2. Zender, C. S., Analysis of Self-describing Gridded Geoscience Data with NetCDF Operators (NCO), Environ. Modell. Softw., Vol. 23(10), P. 1338 – 1342, 2008.
3. Zender, C. S., and H. J. Mangalam, Scaling Properties of Common Statistical Operators for Gridded Datasets, Int. J. High Perform. Comput. Appl., Vol. 21(4), P. 485 – 498, 2007.
4. NCO 4.4.3 User Guide / Интернет-ресурс. – Режим доступа: URL: http://nco.sourceforge.net/nco.html – Загл. с экрана.