ОПТИМИЗАЦИЯ СКОРОСТИ ОБРАБОТКИ ДАННЫХ

Першин Е.А.
Томский политехнический университет
634050, Россия, г. Томск, пр–т Ленина, 30
E-mail: pershinevgen@sibmail.ru

В современном мире основной задачей для любого человека является обработка информации. Информация поступает из различных источников, из интернета, телевидения, и многих других мест. И обработка этой информации требует применения некоторых усилий, и зачастую специальных инструментов для обработки этих данных. Повседневную информацию может обработать один персональный компьютер или ноутбук, но если речь идет о специализированных данных, объем которых занимает десятки и сотни гигабайт, то для этого требуются специализированные вычислительные машины с соответствующим программным обеспечением.

Сегодня скорость обработки информации является очень актуальной проблемой. Информации на различные предприятия поступает все больше и больше, и требуется ее обработать в максимально короткие сроки. В сентябре 2013 года компания Google поставила мировой рекорд по скорости обработке информации. Они обработали один петабайт данных за шесть часов и десять минут. Для этого потребовалось около десяти тысяч компьютеров и более ста тысяч жестких дисков для хранения этой информации. Но не у всех корпораций есть возможность использовать неограниченные ресурсы для обработки данных, поэтому актуальна задача по уменьшению времени обработки данных без изменений в вычислительных мощностях компаний.

Исследование на тему оптимизации скорости обработки информации проводилось на данных формата netCDF. NetCDF (Network Common Data Form) – машиннонезависимый двоичный формат файлов, являющийся стандартом для обмена научными данными. Заголовок формата содержит информацию о содержимом файла. Страница проекта поддерживается программой Unidata объединением университетов в области исследований атмосферы (University Corporation for Atmospheric Research). Формат был создан в 1987 году и используется до сих пор. Первые две версии формата использовались в основном в климатических лабораториях, начиная с версии netCDF-3, выпущенной в 1997 году формат стал использоваться в научных лабораториях различных отраслей для хранения данных. В основном используется в климатологии, например при предсказании погоды, изучении изменения климата и геоинформационных системах. Формат является открытым стандартом.

Набор данных netCDF состоит из трёх основных компонентов. Размерностей (dimensions), переменных (variables) и атрибутов (attributes), каждый из которых имеет имя и идентификационный номер ID . Используя эти компоненты можно описать данные и отношения между ними.

Структура файла:

```
netCDF name {
   dimensions: ...
   variables: ...
   data: ...
}
```

Файлы данного формата создаются по стандарту CORBA. CORBA – это открытый стандарт написания распределенных приложений, который описывает формат данных, которые используются в научной отрасли. Формат netCDF подходит под определение таких данных. В стандарте описываются правила наименования переменных и объектов, правила их следования в структуре.

Первая проблема оптимизации скорости обработки данных данного формата – это несоответствие стандарта и реальных данных. Тысячи научных институтов и лабораторий по всему миру использую формат netCDF для представления файлов, но только небольшая часть данных соответствует стандарту CORBA. Это значительно затрудняет создание технологии для изменения файлов этого формата. В реальных файлах данные могут идти в любом порядке и иметь произвольное название.

Следующая проблема оптимизации скорости обработки данных – это отсутствие инструментов для взаимодействия с данным форматом. Существует пакет инструментов netcdflib для операционных систем семейства Linux. В пакет входят инструменты для просмотра данных, просмотра заголовков, и интерфейсы для языков программирования для взаимодействия с данными файлами. Пакет поддерживает такие языки как C, C++, Java, IDL, Fortran, Python, Perl, MATLAB. Но данный пакет дает очень ограниченные возможности для редактирования файлов. Задачи вроде упорядочивания данных или изменения порядка переменных невозможно решить, используя только данный пакет инструментов. Unidata занимаются развитием формата данных и данного пакета инструментов, но за период с 2004 по 2013 год в данном SDK не было сильных изменений.

В поисках способа оптимизации при работе с netCDF было проведено исследование на тему скорости открытия файлов в зависимости от объема. Было предложено изучить это, и, в случае положительных результатов исследования, создать программу по объединение нескольких netCDF файлов в один.

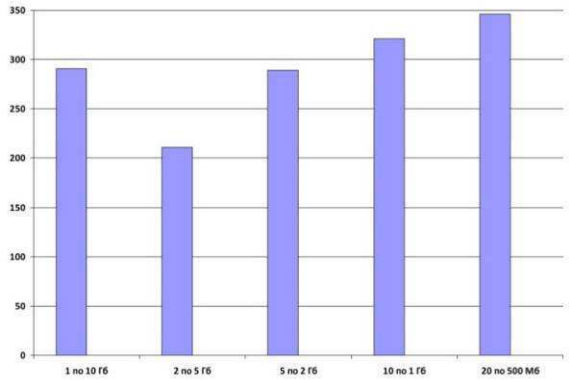

Рис. 1. Результаты эксперимента #1

Для проведения эксперимента была создана программа, которая открывает netCDF файл, считывает заголовок и закрывает этот файл. Измерение скорости работы программы измерялось с помощью утилиты VTune. VTune – это программа для измерения производительности программ. В рамках исследования было проведено пять опытов: один файл размером десять гигабайт, два файла размером пять гигабайт, пять файлов размером по два гигабайта, десять файлов размерами по одному гигабайту и двадцать файлов размерами по пятьсот мегабайт. На рисунке 1 показаны результаты исследования. Погрешность измерений с общим объемом 10 гигабайт примерно 10 миллисекунд. Лучший результат показал эксперимент, в котором открывалось пять файлов по два гигабайта. Файлы открылись всего за 211 миллисекунд. Следующие результаты, это опыт с пятью файлами, закончивший работу за 289 миллисекунд. Практически за такое же время открылся один файл в десять гигабайт – за 291 миллисекунду. Худшие результаты показали оставшиеся два эксперимента, показавшие результаты в 321 и 346 миллисекунд.

Исходя из результатов эксперимента, можно сделать выводы, что существует файлы некоторого объема данных, которые открываются быстрее всего. При увеличении данного объема, производительность открытия файла падает, так же как и при многочисленном разбиении данных на разные файлы.

Следующий эксперимент позволит определить примерный объем данных, который открывается быстрее всего.

Был проведен эксперимент, аналогичный первому, но с другим набором данных. Был увеличин максимальный объем данных и взято разбиение от одного до пяти гигабайт с шагом в двести мегабайт. Погрешность измерений увеличилась до 20 миллисекунд. Лучшие результаты показали эксперименты, открывающие файлы от 2,4 гигабайт до 3,8 гигабайт. Разница между этими экспериментами незначительная, входящая в диапазон погрешности в 20 миллисекунд. Остальные эксперименты показали худшие результаты.

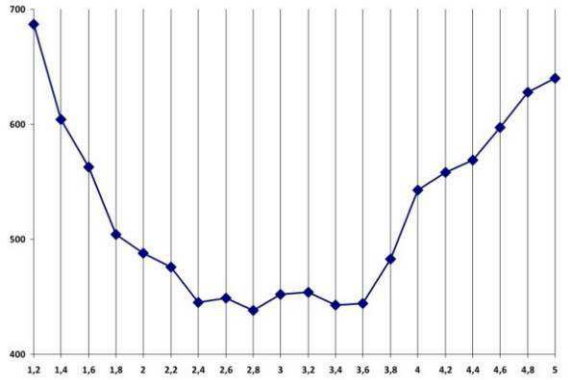

Рис. 2. Результаты эксперимента #2

По результатам второго эксперимента, которые можно увидеть на рисунке 2, можно сделать вывод, что быстрее всего открываются файлы netCDF объемом от двух с половиной до трех с половиной гигабайт. В отношении небольшого количества файлов это не очень большая величина, которая может казаться незначительной, но в институтах, где используется данный формат файлов, обрабатываются тысячи файлов с общими объемами сотни и тысячи гигабайт, где разница в 100 миллисекунд переходит в разницу, измеряемую в минутах.

Для небольшой оптимизации обработки netCDF файлов, можно несколько небольших файлов с одинаковой структурой объединять в файлы по три гигабайта, и обрабатывать данные из них. Но преобразование файлов так же занимает некоторое время и ресурсы, поэтому данная оптимизация реализуема только для систем, которые обрабатывают данные в реальном времени и им важна каждая минута. Если данные обрабатываются в свободном режиме, то данная оптимизация не целесообразна, потому что относительно времени обработки самих данных, время открытия незначительно мало.

Литература

1. NetCDF // Википедия. [2013—2013]. Дата обновления: 13.03.2013. URL: http://ru.wikipedia.org/?oldid=53431010

2. UNIDATA // UNIDATA [2013]. Дата обновления: 2.08.2013. URL: http://www.unidata.ucar.edu/

3. Формат netCDF // Океанология. Океанография [2013]. Дата обновления: 14.05.2013. URL: http://www.oceanographers.ru/forum/viewtopic.php?t=80

4. OMG CORBA // CORBA Web Site [1997-2013]. Дата обновления: 19.08.2013. URL: www.corba.org/